People detection, tracking and re-identification through a video camera network

Malik Souded

UNIVERSITÉ DE NICE - SOPHIA ANTIPOLIS

ÉCOLE DOCTORALE STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

T H È S E

pour l'obtention du grade de

Docteur en Sciences

de l'Université de Nice - Sophia Antipolis

Mention : AUTOMATIQUE TRAITEMENT DU SIGNAL ET DES IMAGES

présentée et soutenue par

Malik SOUDED

PEOPLE DETECTION, TRACKING AND RE-IDENTIFICATION THROUGH A VIDEO CAMERA NETWORK

Thèse dirigée par François BRÉMOND

Soutenance prévue le 20/12/2013

Jury:

Monique	THONNAT	Directrice, INRIA Sophia-Antipolis, France	Présidente
James	FERRYMAN	Professeur, University of Reading, UK	Rapporteur
Carlo	REGAZZONI	Professeur, University of Genova, Italy	Rapporteur
Patrick	BOUTHEMY	Directeur, INRIA Rennes, France	Examinateur
François	BREMOND	Directeur, INRIA Sophia-Antipolis, France	Directeur de thèse
Marie-Claude	FRASSON	Directrice, Digital Barriers, Sophia-Antipolis, France	Invitée

Thèse

Dection, Suivi et Re-identification de Personnes à Travers un Réseau de Caméra Vidéo

People Dection, Tracking And Re-identification Through a Video Camera Network

Malik Souded

Le 20/12/2013

RÉSUMÉ

Cette thèse a été effectuée dans un contexte industriel et présente un framework complet pour la detection et le suivi de personnes dans un réseau de caméras de surveillance. Les trois principales étapes du processus sont traitées: la détection de personnes, le suivi de personnes dans un contexte mono-caméra et enfin la ré-identification de personnes dans le contexte multi-caméras. Les performances élevées, la généricité et la facilité de déploiement ainsi que le traitement en temps réel sont les contraintes fortes qui ont guidé ces travaux. Certaines parties du travail proposé ont déjà été intégrées et déployées dans un produit commercial de vidéo surveillance intelligente alors que les autres parties sont à l'état de prototypes et seront intégrés dans un futur proche.

La détection de personnes vise à localiser et délimiter les personnes sur les séquences vidéo ainsi que sur les images statiques. Le détecteur de personnes proposé appartient à la catégorie des détecteurs de silouhette entière et opère à l'aide d'une cascade de classifieurs, appris en utilisant l'algorithme LogitBoost sur les descripteurs de covariances de régions. Une approche de l'état de l'art, fournissant de bonnes performances mais non applicable pour le traitement en temps réel a été prise comme base de travail et a été optimisée afin de permettre le traitement en temps réel tout en améliorant légérement les performances de détection. La méthode d'optimisation proposée est généralisable à de nombreux autres types de détecteurs basés sur les cascades de classifieurs, et dont l'espace de tous les classifieurs faibles possible ne peut Ãªtre testé exhaustivement dans un temps raisonable.

Le suivi de personnes dans le contexte mono-caméra vise à fournir un ensemble d'imagette de chaque personne observé par chaque caméra, afin de permettre le calcul de la signature visuelle de ces personnes. Il fournit aussi certaines informations du monde réel qui sont trÃ¨s utiles pour améliorer les résultats de la ré-identification, du moment que ce suivi est réalisé en utilisant des caméras statiques et calibrées. Ce suivi de personne est effectué à l'aide du suivi de points d'intérêt SIFT en utilisant un filtre à particule spécifique, prenant en compte un certain nombre d'informations utiles telles que le résultats de la soustraction de fond et la mesure de fiabilité des descripteurs SIFT que nous proposons dans ce travail, en plus d'un framework d'association de données qui permet d'inferer le suivi d'objets à partir du suivi des points SIFT, et qui permet de gérer la plus part des cas possibles, particulièrement les occultations.

Enfin, la ré-identification de personnes est effectuée à l'aide d'une approche de type apparence globale. Une approche de l'état de l'art, permettant un traitement en temps réel mais fournissant des performances très variables en fonction des données fournies en entrée, est améliorée afin de fournir de meilleures performances tout en maintenant l'avantage du traitement en temps réel. Les ameliorations ont été introduites à différents niveau du traitement de l'approche originale, soit en remplaçant certaines étapes initiales ou en ajoutant de nouvelles. Une partie "connaissance du contexte" a été introduite afin de rendre la signature visuelle plus robust aux changement d'orientation des personnes, assurant de meilleures performances de ré-identification dans le cas d'applications réelles.

Cette thèse fourni les contributions suivantes: Un détecteur de personnes qui propose (1) une approche généralisable de clustering pour les données néganives avant l'apprentissage du détecteur, accélérant la phase d'apprentissage et optimisant le détecteur dont le traitement devient plus rapide (temps réel) et plus précis (de meilleurs performances). Un framework de suivi d'objets, basé sur le suivi de points d'intérêts SIFT à l'aide d'un filtre à particules, en plus d'un processus d'association de données, qui propose: (2) une méthode de détection et de sélection d'un nombre constant et correctement réparti de points SIFT sur l'objet d'intérêt, permettant une meilleure représentation de l'objet et de ce fait, de meileures performances de suivi particulièrement dans le cas d'occultations partielles, (3) une méthode hybride de pondération de particules, qui améliore le suivi des points SIFT, prenant en compte la mesure de similarité du descritpeur SIFT ainsi que les résultats de la soustraction de fond d'une manière plus complexe

qu'une simpe pondération binaire, (4) une méthode d'association de données détéctant toutes les situations possibles durant le suivi (incluant les occultations partielles et complètes) et traitant chancune d'elles. Cette méthode d'association de données utilise la position des points SIFT et leurs mesure de fiabilité (introduite dans l'étape précédente) pour identifier l'état de chaque objet détecté/suivi et de mettre à jour les états de l'ensemble des objets suivis, (5) une methode rapide (temps réelle) de gestion des occultations, utilisant les points SIFT suivis, l'information couleur, ainsi que certaines données du monde réel (véritable dimensions, véritable vitesse), apprises durat le suivi de l'objet, afin de ré-acquérir l'objet occulté lorsqu'il réapparait. Enfin, pour la détection de personnes, une méthode de l'état de l'art is fortement améliorée avec (6) une máthode temps réel pour l'allignement des images d'une même personne pour minimiser les erreurs de detection des personnes (7) l'enrichissement de la signatuer visuelle en ajoutant l'information de la texture sous la forme de descripteurs SIFT et de matrices de covariance encodant la couleur et la texture en même temps pour la description des patches RHSP, (8) la classification de la face visible de chaque personne sur chaque image, permettant de calculer des signatures visuelles pour chaque classe, augmentant l'efficacité de ces signatures visuelles et permettant aussi une meilleure pondération de chaque type d'information utilisée (9) l'utilisation de l'information fourni par la calibration des cameras et le suivi mono-camera des personnes pour filtrer les candidats dont l'état ne respecte pas les contraintes spatio-temporelles (10) une method de ponderation automatique et adaptative pour mieux focaliser l'algorithme de re-identification sur l'information la plus discriminante, et de diminuler ou supprimer l'importance de certain descripteurs locaux.

ABSTRACT

This thesis is performed in industrial context and presents a whole framework for people detection and tracking in a camera network. The three main processing steps are addressed: people detection, people tracking in mono-camera context, and people re-identification in multi-camera context. High performances, system genericity and ease of deployment, and the real-time processing are the most important constraints which have guided this work. Some parts of the proposed work are already integrated and deployed in a commercial product while others are in prototype state and are planned to be integrated in future.

People detection aims to localise and delimits people in video sequences and static images. The proposed people detection is a full body one and it is performed using a cascade of classifiers trained using LogitBoost algorithm on region covariance descriptors. A state of the art approach, providing good performances but not applicable for real time is taken as basis and is optimized to process in real time while detection performances are slightly improved. The optimization scheme is generalizable to many other kind of detectors based on cascade of classifiers where all possible weak classifiers cannot be reasonably tested.

People tracking in mono-camera context aims to provide a set of reliable images of every observed person by each camera, to extract his visual signature for re-identification purpose. It provides also some real world information which are useful to improve re-identification process, as long as this mono-camera tracking is performed using static and calibrated cameras. It is achieved by tracking SIFT features using a specific particle filter, taking in account many useful information like background subtraction results and a proposed reliability measure of SIFT descriptors, in addition to a data association framework which infer object tracking from SIFT points one, and which deals with most of possible cases, especially occlusions.

Finally, people re-identification is performed using an appearance based approach. A state of the art approach, which performs in real time, but provides various performances depending on the input data is improved to provide better performances while keeping the real-time processing advantage. The improvements are introduced at different levels of the original approach, by replacing some of initial steps or by adding new ones. A context-aware part is introduced to robustify the extracted visual signature against people orientations, ensuring better re-identification performances in real application case.

This thesis makes the following contributions: A people detector which proposes (1) a generalizable clustering approach for negative data before people detector training, speeding-up training process and optimizing the trained detector which performs faster (real-time) and better (performance improvements). An object tracking framework, based on SIFT feature tracking by particle filter and data association process, which proposes: (2) a method to detect and select constant number of well distributed SIFT points on the object of interest for tracking, allowing better representation of the object and thereby, a better tracking performances especially in partial occlusion situations, (3) an hybrid particle weighting method, which improves SIFT points tracking, taking in account the SIFT descriptor similarity measure and the background subtraction result in a sophisticated way (not a simple binary weighting), (4) a data association process to detect all possible situations during tracking (including partial/full occlusions) and to manage each of them. This data association process use the tracked SIFT points localisation and their reliability measures (introduced in the previous step) to identify the state of each detected/tracked object, and to update the whole tracked object states, (5) a fast (real-time) occlusion management method, using tracked SIFT points, color information, and some other "real world" information (real dimensions, real velocity), learned during object tracking, to reacquire occluded objects after their reappearance. Finally, for people re-identification, a state of the art method is strongly improved by (6) a fast method for images alignment for multiple-shot case, to reduce people delimitation error in images and allow same parts comparison (7) the add of texture information to the computed visual signatures, by adding SIFT features as a new

feature in the signature and by characterizing RHSP patchs by covariance descriptors encoding both color and texture information at the same time, (8) a method for people visible side classification, allowing to compute more accurate and discriminant visual signatures for each class, and allowing a better feature weighing (9) a method to use camera calibration information to filter candidate people who does not match spatio-temporal constraints (10) an adaptive feature weighting method to allow each re-identification query to focus on the more discriminant features, and to reduce or cancel local feature weights in some cases, according to visible side classification.

ACKNOWLEDGMENTS

First, I would like to express my gratitude to my thesis supervisor, François BRÉMOND for the quality of the supervision, for helping me to improve my research skills, for guiding me while allowing me the necessary freedom to explore different directions. I would like to thank him also for having well managed the industrial context of the thesis while ensuring the research excellence, and for being patient with me, having good as well as bad work periods.

I would like to greatly thank Carlo REGAZZONI, professor at Genoa university, and James FERRYMAN, professor at Reading university, for accepting to review my PhD manuscript and to be part of the committee. Their pertinent feedbacks have been enriching and constructive and will help me to improve my future work.

I would like to thank Monique THONNAT for accepting to be the president of the committee. Her many advices on my manuscript writing and for my defence preparation have been very useful. I would also like to thank Patrick BOUTHEMY, research director at INRIA Rennes for accepting to be my thesis examiner.

I would like to thank Digital Barriers company (previously Keeneo when I started my PhD) for hosting me and allowing me to perform this work. I would like to particularly thank Benoit GEORIS, ex-Director of Keeneo who opened me the doors of the company, Alberto AVANZI without whom this adventure would not have happened and who provided me with invaluable help in both technical and daily life aspects. I would also like to thank Marie-Claude FRASSON who was my manager in the company. I thank her warmly for her effective management, balancing performance requirement and the continuous search for good and warm working conditions in the team. Special thanks to Robert STAHR also for being always available and attentive, for his valuable technical assistance and also for being the big brother when I had to go through difficult periods. I would like to thank Laurent GIULIERI for his non-negligible help during my first year of PhD. I also thank all other colleagues and friends I met in the company, Kjetil JACOBSEN with whom I was jogging to clear my mind, Guilllaume BARRELET we are missing, me and the Green King, Audrey MOLITISANTI, Olivier LEROUX, Sebastien WEBO and Dario RAVARRO for their friendship and availability and finally, Thomas HERLIN, another big brother who has given me many good advices.

Many special thanks are also to all of my colleagues in the STARS team for their kindness as well as their scientific and technical supports during my thesis period, especially Phu and Slawomir. Thanks to both Julien(s) for the pleasant moments playing table football and sharing some common hobbies. All of them have brought a lot of warmness and friendship, and have made this PhD period more pleasant.

I would like to thank also Cécile PHAM-VAN at Digital Barriers and Jane DESPLANQUES at INRIA for their invaluable help in the management of administrative aspect of every day, allowing me to focus on my research work.

I finish these acknowledgements by the people who matter the most to me in this world. I thank my parents for their encouragement and support, without which it would have been impossible to finish. I thank and dedicate this work to my two brothers Yacine and Walid and to my sister Yasmina. Finally, I conclude this acknowledgements with my fiancée Amel who has always been behind me, pushing me forward, and particularly constantly put up with me, because I have not always been easy to live with during this period. Thank you all.

I would like to thank and to present my excuses to all the persons I have forgotten to mention in this section.

Malik SOUDED

Sophia Antipolis (France), December 2013

CONTENTS

FIGURES

TABLES

1

INTRODUCTION

1.1 Motivation

With the rapid technological advances of the last 15 years and the easy accessibility of cameras and digital media storage, video surveillance is widely used and developed in all aspects of life in modern societies. The recent global security context, with the terrorist attacks in New York (2001), Madrid (2004) and London (2005) have obviously contributed to this growth, to ensure the protection of people and assets against terrorism acts, but not only. The recent attack in Boston (2013) demonstrates the effectiveness of video surveillance, thanks to which the two authors were quickly identified. The prevention and repression of crime and delinquency, the protection of industrial and administrative buildings, securing airports, train stations and ports, road safety, people flow management and other needs have shown the necessity to increase the capacity of video surveillance of cities, businesses, and other concerned stakeholders.

Video surveillance systems are used both in live and differed modes. Live streams are viewed by video operators in real time, allowing rapid interventions if an event of interest is detected. But the high number of deployed cameras and the volume of incoming data hinder the live processing of all streams. For these reasons, many acquired streams are stored on digital media for a predefined time, and are used a posteriori to replay any event of interest and extract useful information, like person identification or evidence extraction.

This continuous increase in the number of deployed surveillance cameras increases the workload of video operators in both modes.

Figure 1.1: Video-surveillance center: too many screens and difficulty to monitor all screens.

In live surveillance mode, each video operator has to monitor more streams, causing a decrease in efficiency. The large number of video streams assigned to each video operator increases the probability of missing important events when operator is not focussing on the right camera at the right time, because he/she is monitoring many screens, or by the fact that one screen is dedicated to many cameras and displays the viewed scene of each of them periodically. In addition to that, some recent medical studies have shown that after 20 minutes of focusing on surveillance screens, a video operator lose 90 percent of his concentration and vigilance. Most of the time, no particular event occurs so after a while of monitoring images where nothing happens, sleepiness of vigilance, fatigue and decreased concentration often cause failure in some important event detection.

For the a posteriori mode, especially for government security agencies like police, searching for a given person of interest in hundreds or thousands hours of recorded videos, provided by many cameras, requires to assign a large number of enforcement officers to this task, and requires a lot of hours or days to be performed. Dedicate so much manpower is often complicated, due to the high financial cost and the unavailability of competent agents who are assigned to other tasks. In addition to that, enforcement officers are humans and thereby, they are subject to the same problems of fatigue and decreased concentration similarly to video operators.

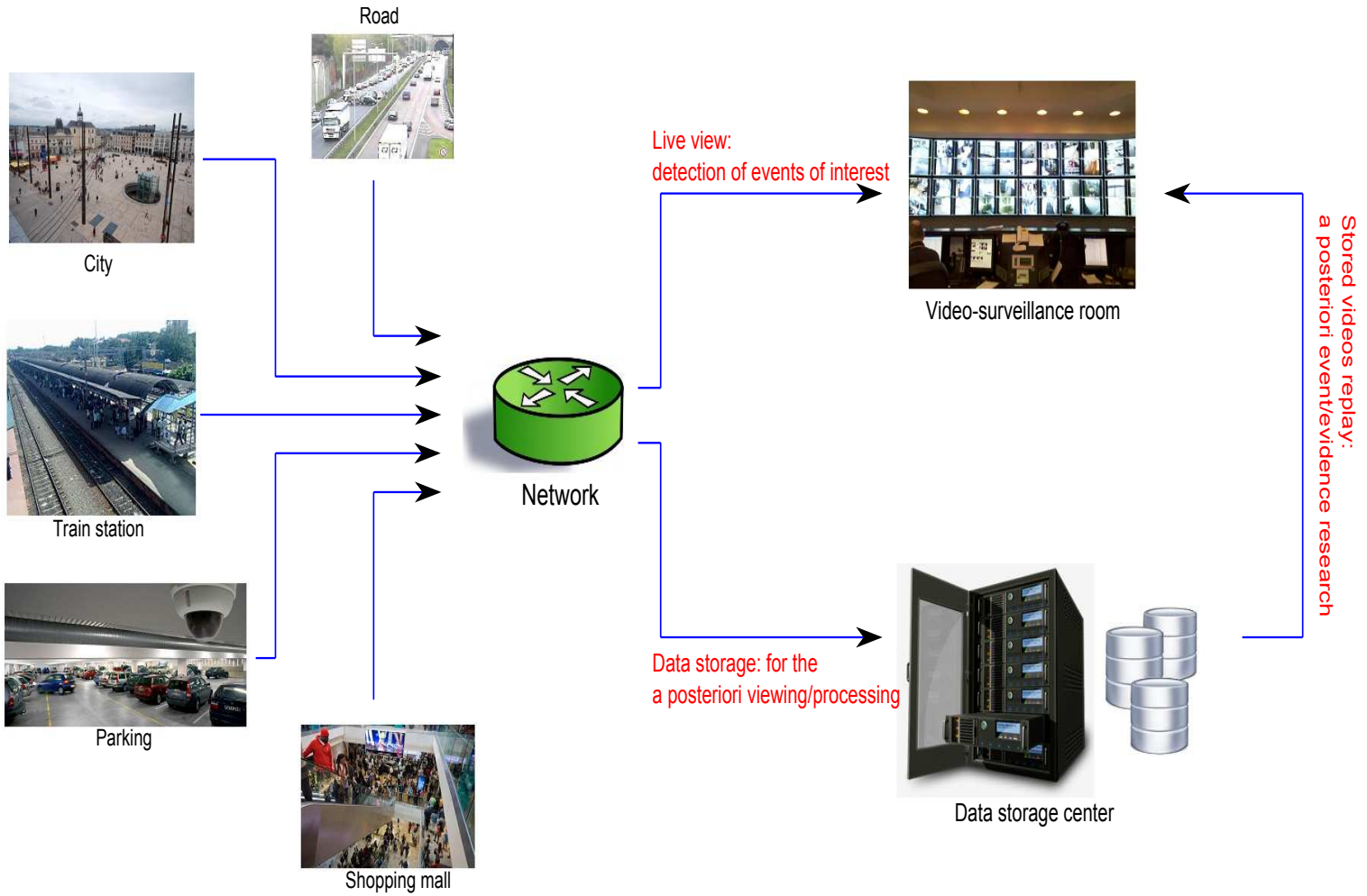


Figure 1.2: Video-surveillance architecture: live viewing and a posteriori viewing.

To overcome all these problems and process all the video data provided as live streams and recorded videos, there is a growing demand of automated analysis and understanding of video contents. Reproducing human analysis and reasoning on observed events is becoming a critical field of research. This research domain covers many tasks like object detection and recognition, tracking in mono-camera and more recently in multi-camera contexts, gesture recognition, behaviour analysis and understanding, etc. All these tasks are used in many domains like robotics, entertainment, but also and in a large proportion, in security and video surveillance.

This thesis takes place in this context and consists in answering to this question: How an automatic system can characterize a person of interest, to track him/her in real time in a camera network and determine his/her localization in a huge volume of recorded videos?

1.2 Context of Study

My work was conducted under a collaboration between STARS team from INRIA and Digital Barriers company.

STARS (Spatio-Temporal Activity Recognition Systems) team is an INRIA research team, which focuses on the design of cognitive vision systems for Activity Recognition. The research team is interested in the real-time semantic interpretation of dynamic scenes observed by video cameras and other sensors. It studies long-term spatio-temporal activities performed by agents such as human beings, animals or vehicles in the physical world. The major issue in semantic interpretation of dynamic scenes is to bridge the gap between the subjective interpretation of data and the objective measures provided by sensors. To address this problem Stars develops new techniques in the field of cognitive vision and cognitive systems for physical object detection, activity understanding, activity learning, vision system design and evaluation. Stars focuses on two principal application domains: visual surveillance and healthcare monitoring. STARS has two main research themes: Scene understanding for activity recognition and Software architecture for activity recognition.

Digital Barriers is a security company which provides advanced surveillance technologies to the international homeland security and defence markets. It brings innovative thinking and solutions to the protection of most critical national assets, locations and infrastructure. It combines a long heritage in the security and defence sectors, with operational expertise and an understanding of how best to apply and deploy emerging technologies. Digital Barriers is specialised in delivering intelligent surveillance information from challenging environments. It conducts advanced research in computer vision

for video surveillance purpose. Among the issues that the company is interested, we can mention securing sensitive sites against all intrusion, recognizing suspicious behaviours like loitering, detection of dangerous events like abandoned luggage in public places and tracking the person who has abandoned it across the camera network which covers this public place. It is also interested on forensics fields, by finding evidences and people a posteriori in recorded videos.

The work presented in this thesis is also directly related to an ITEA2 European project called **ViCoMo** (Visual Context Modelling), in which both STARS and Digital Barriers have participated. The ViCoMo project was a 3 years project, started in September 2009 and ended in November 2012. The ViCoMo project is developing advanced video-interpretation algorithms to enhance images acquired with multiple camera systems. By modelling the context in which such systems are used, ViCoMo significantly improves the intelligence of visual systems and enables recognition of the behaviour of persons, objects and events in a 3D view. The project enables advanced content and context based applications in surveillance and security, and transport/logistics with spin-offs in the consumer & multimedia domains.

1.3 Issues in People Detection, Mono-Camera Tracking and Person Re-identification

1.3.1 From Practical Point of View: Large scale video-surveillance constraints

As mentioned before, the aim of this thesis is to provide an automatic system to re-identify people across a camera network (see figure 1.3). Many studies have been done on this topic, using various techniques and under different constraints. We can divide these approaches into two main types: biometric and appearance-based approaches.

The main types of biometric techniques for re-identification in video surveillance context are face recognition and gait recognition. We can mention also iris and fingerprint analysis as other biometric techniques, but these last ones can not be considered as a mean of (re)identification for wide scale video surveillance field, because they require a cooperation and voluntary actions from individuals evolving in the monitored environment. Indeed, capturing the iris image for analysis requires the person of interest to position his eye close to a specific sensor and directly in front of it (See figure 1.4). Similarly, the capture of fingerprint of the person of interest requires from him to put his finger on a specific sensor (See figure 1.5). At the opposite, face recognition (Figure 1.6) and gait recognition (Figure 1.7) techniques do not require (or in a lesser extent for face

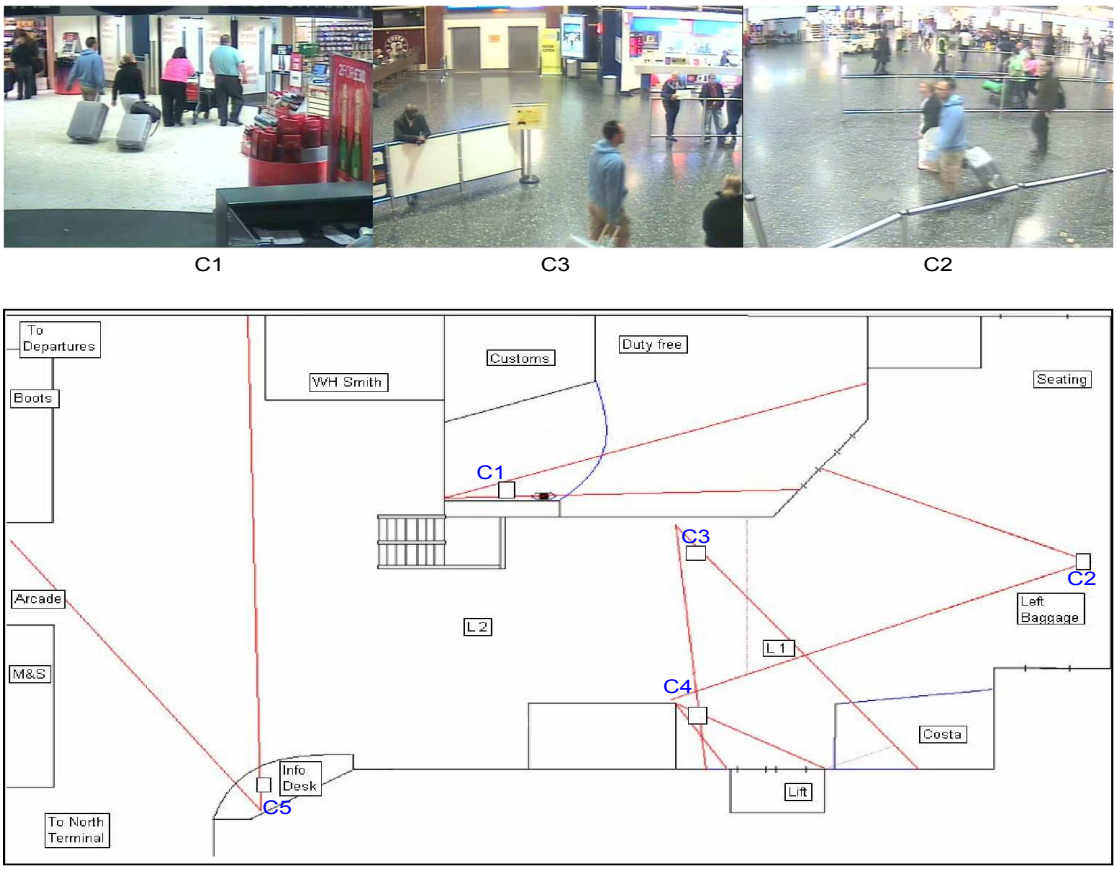


Figure 1.3: iLids multi-camera tracking challenge environment: 5 cameras with and without overlapping of field of views. C1, C2 and C3 are the viewed scene of the corresponding cameras. The position and orientation of each camera is displayed in the floor plan.

recognition) the collaboration of observed people. They can be applied in large areas with many people and many entrance and exit possibilities as long as few constraints are verified.

Many approaches have been proposed for face recognition [Bauml 2010, Belhumeur 1997, Kirby 1990, Lee 2003a] and gait recognition [Wang 2003b, Lee 2003b, Chellappa 2007]. When the conditions and constraints are satisfied, some of these approaches provide high performance, especially for face recognition. Unfortunately, some of these constraints are not satisfied by most deployed surveillance systems.

First, for economic and optimization reasons, most of surveillance cameras cover large fields of view, providing images with small persons/objects of interest, so the extractable information is insufficient to provide acceptable performance after processing

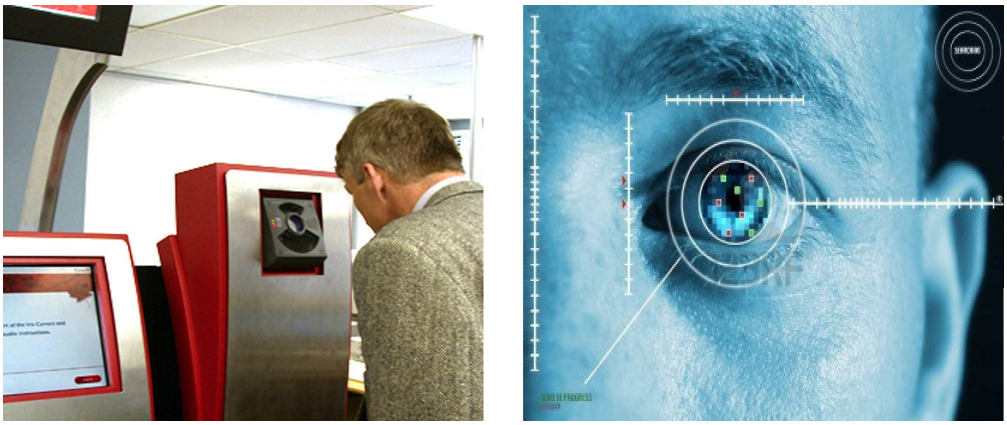


Figure 1.4: Iris identification: Collaboration and specific actions from people are required. The person has to put his eye in front of the sensor.

Figure 1.5: Fingerprint identification: Collaboration and specific actions from people are required. The person has to put his finger on the sensor.

(face size under the minimum required size, indistinguishable gait, etc.).

Second, the large number of deployed cameras in a given site and the distance between this site from the monitoring or storage location may overload transfer network. For this reason, the streams are generally compressed by the cameras, sent to the processing or storage location, before to be decompressed for display or for processing. The storage of huge volume of videos also requires compression. Most of compression/decompression techniques introduce a loss of information and some noise. These issues directly impact the performance of face and gait recognition.

Finally, biometric approaches are significantly dependent on the point of view of the camera and the orientation of the person of interest with respect to this camera. If the face is not visible (the person is seen from behind or from the side), face recognition

Figure 1.6: Face recognition: People in front of camera (green rectangles) can be processed for face recognition. The person who is seen from behind (red rectangle) cannot be processed. The last person seen from the side (orange rectangle) is in the limit of constraints and processing can fail. Here, collaboration of people is not required.

cannot be performed.

For these reasons, we have oriented our work on appearance-based approaches, which have less constraints than biometric ones, and are more adapted to the video surveillance requirements.

Achieving people re-identification with good performances requires to use reliable information as input. For this reason, this thesis focuses on three main parts. The first part consists in people detection on static images and video sequences. The second part consists in people tracking in mono-camera context. The last task, which is the final aim of this thesis, consists in people visual-signature extraction and comparison for re-identification.

These three parts are interlinked and the performances of any of them affect directly one or both other parts. The dependence of these three parts is shown in the figure 1.8.

People detection on static images and video sequences provides/validates the targets to track to mono-camera tracking algorithm in collaboration with a background subtraction algorithm (this collaboration will be detailed in chapter 5) and provides the candidates for re-identification to the re-identification algorithm.

Mono-camera tracking algorithm provides the different locations and images of a given person through out time to the re-identification algorithm, allowing it to learn a robust visual signature for this person, taking into account observed variations of his/her images.

In the same time, re-identification algorithm allows mono-camera tracking algorithm to deal with occlusions, by maintaining tracks of partially occluded objects thanks to

Figure 1.7: Gait recognition: gait is decomposed into different identifiable phases which are processed. No collaboration from people is required.

their partial visual signatures and by reacquiring lost targets using their whole visual signatures.

Two important points can be noted:

- The distinction between static images and video sequences in people detection, which will be more detailed in the corresponding chapter, is mainly due to the features used in each case and to the type of addressed problems.

- Even if the final aim of this work is mainly dedicated to people detection, mono-camera tracking and re-identification on video sequences, the second part concerning mono-camera will be generalized to all types of objects of interest, and not only people. This choice is motivated by the fact that object tracking is an important task in many security applications. Digital Barriers is interested by this generic object tracking algorithm for other purposes (for example: car tracking). Never-

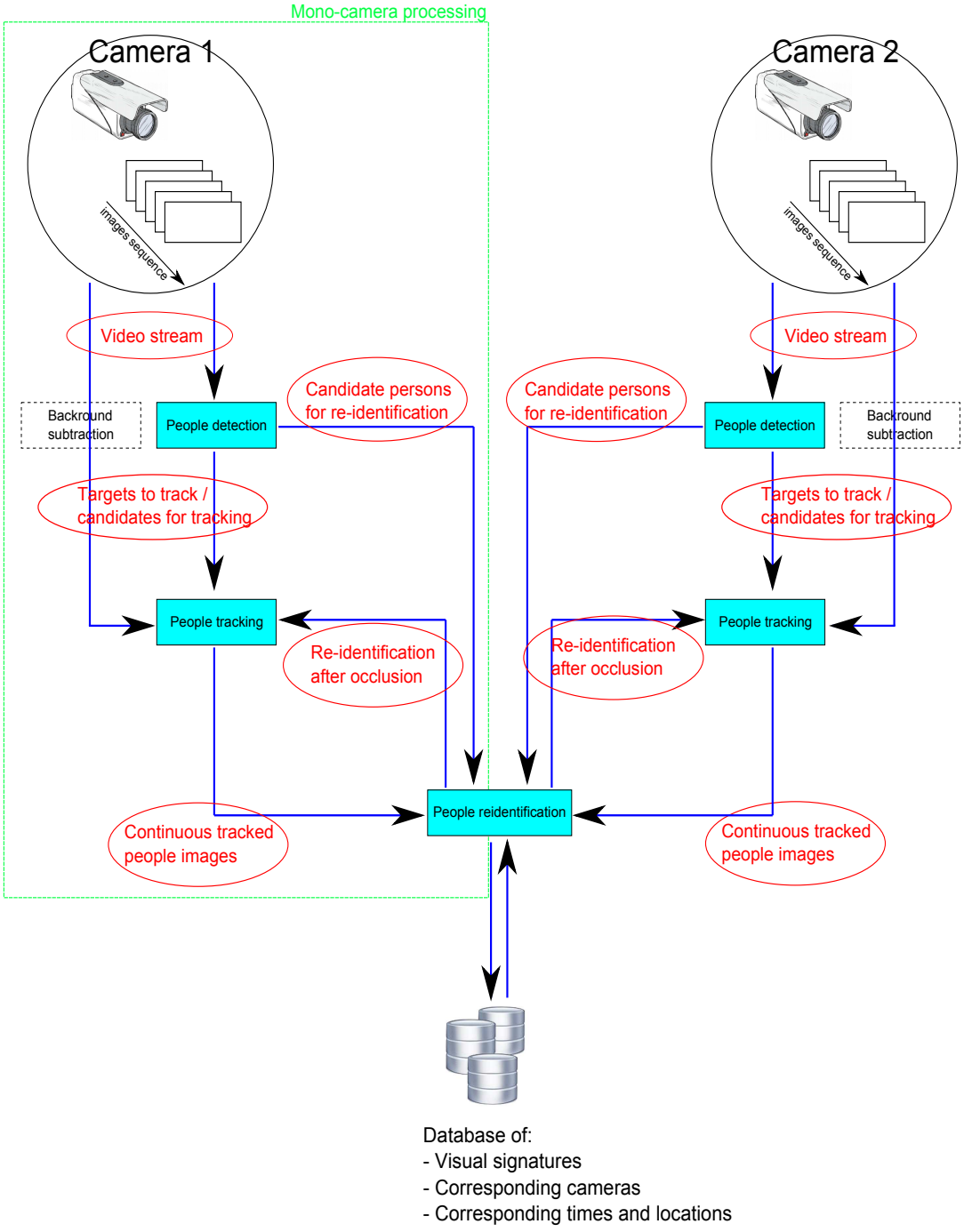


Figure 1.8: Dependency between people detection, mono-camera tracking and re-identification. In this figure, two cameras are used but the architecture can be generalised for a network of many cameras.

theless, we will take care to report and discuss the specificities of people tracking whenever it is necessary.

1.3.2 From Computer Vision Limitations Point of View

Automated person detection finds its applications in many areas including human-robot interaction, surveillance, pedestrian protection systems, automated image and video indexing. In the video surveillance context, detecting the whereabouts of humans is the first requirement if any event involving people has to be detected. This ranges from simple events like intrusion in forbidden area to complex events like behaviour analysis.

Object tracking is the task of following one or more objects in a scene, from their first appearance to their exit [Forsyth 2002]. An object may be anything of interest within the scene that can be detected and depends on the requirements of the application. Given a sequence of image frames to track a set of objects, these objects correspond to sub-images, in each frame. In general, in a dynamic environment both background and objects may vary. In principle, solving this general unconstrained problem is hard. One can add a set of constraints to help solving this problem. The more the constraints, the easier the problem is to solve.

Person re-identification consists in determining if a person of interest has already been observed over a network of cameras (see figure 1.3). As mentioned before, this task can be performed using different approaches depending on constraints and available information, using biometric-based techniques like face and gait recognition or using appearance-based approaches. Due to the cited reasons before, concerning the difficulty to ensure biometric approach constraints, appearance-based approaches are more suitable for large wide video-surveillance purpose. The global appearance of an individual has to be modelled to re-identify it. This modelling is the last topic of our work.

All these tasks are complex to achieve, due to the numerous challenges which affect their performances. Some of these challenges are common to the three tasks, and affect them with different degrees, and some other challenges are specific to a given task. In the following paragraphs, we present the most important challenges and we discuss how they affect our tasks.

- **Physical variation of persons**: Person appearance can vary greatly. People do not share a common body size, color, texture, and the appearance is highly influenced by the clothing they wear. This is a specific issue for people detection only, and can be considered as an advantage for mono-camera tracking and re-identification

tasks.

In fact, People detection consists in finding the most common characteristics of all people which are absent in other object classes. This large variations in physical aspect requires to "enlarge" the model of persons and thereby, increase the risk to have false positives (objects which are not people but detected as such because their appearance is close to the one of a person). In the same time, trying to avoid false positives by restricting the person model may cause miss detection of persons with physical appearance out of the bounds of the model. This issue is an important but not a critical one. Most of approaches ignore color and textures, and are based on external shape or part detection in gray scale level domain. The size variation is managed by a multi-scale detection.

Concerning mono-camera object tracking and re-identification, this physical variation can be considered as an advantage due to the nature of this two tasks. unlike people detection which consists in separating a class of objects from all others, mono-camera object tracking and people re-identification require to separate an individual entity from all other entities, from the same class or not. The more physical variation are, the more likely to better characterize an individual entity is.

- **Body deformations**: For a task that depends highly on the shape of a person, body shape deformations can adversely affect the performances. This issue impacts the three tasks, but at different degrees.

  For people detection, it is a critical issue. It is hard to build a unique model for all possible body shapes. Body deformation introduces a variation in parts localization/visibility. A people detector which is build to detect a standing person will probably fail to detect a crouching or a sitting person. A people detector based on part detection can have more chances to detect than a global person detector, but it can also fail if some important parts are not detected (legs of crouched person are hard to identify). The most used solution for this issue is to build as many detectors as identified body shapes. This solution implies a very important detection time, due to the necessity to use all detectors without any a priori information. Another issue of this solution is that some intermediate states of the body shape can be miss-detected. In fact, to pass from the standing state to the crouching one, the body part algorithm acquires many intermediate states which can not be included in existing models.

  For appearance-base people re-identification also, this issue is an important one. Most of appearance models are built by taking into account the localisation of each interesting feature on the person body. A learned visual signature of a standing

person in a camera can be useless if this person appears in another camera in different position (sitting for example). Some interest features can be observed in different position than extracted/learned ones, or can not be visible.

For mono-camera tracking task, this issue can be relatively well managed in most of cases. If the video frame-rate is sufficient, the body deformation occurs continuously, and the whole deformation can be divided into small successive deformations, which can be recognized and managed. With a correct similarity measure of the used features for tracking, successive states of deformation can be linked and the tracked target identity can be maintained.

- **Illumination**: For a task that depends on the lighting conditions, varying illuminations and shadings in different environments can affect the performances.

  For people detection, this issue is a medium one. Depending on the used features for people modelling and the use of illumination normalization step, most of state of the art approaches deal more or less well with this issue.

  For mono-camera object tracking, this issue is more or less important, especially if the light variation occurs suddenly. Depending on the robustness to light changes of the used features for tracking, the similarity measures can greatly vary and cause tracking failure. Nevertheless, the temporal continuity of the processing with a sufficient frame-rate and the localisation of tracked objects before and after illumination changes provide a useful way to decrease failure probability.

  For people re-identification, this issue is a critical one, especially in the case of cameras without overlapping field of view. The same person observed with two cameras under different illumination conditions can have two different appearances. This phenomenon is extreme in the case of bad light acquisition. Unfortunately, the temporal continuity of mono-camera object tracking is missing here, increasing the importance of this issue.

- **Viewpoint changes**: Depending on which angle people/objects are viewed, different shapes can be observed with varying aspect ratios.

  For people detection task, this issue is similar to “Physical variation of persons” one. The variation in viewpoint implies a variation in appearance and size of persons in addition to aspect ratio changes also. For the reasons cited in the “Physical variation of persons” issue, this issue is a critical one.

  For mono-camera people tracking, this issue is a lower one. In the case of static cameras, this issue does not exist. In the case of moving cameras, if the motion is uniform or if the frame rate is sufficient, the temporal continuity of information

allows to deal with successive small variations, and thereby, maintain the track objects identities.

For people re-identification, this issue is more important. The observed features in a given camera can not be visible in another camera, or can be observed with different size and aspect ratio, causing the failure of the re-identification.

- **Crowded scene**: The number of persons in the scene is an important parameter for the task performances.

  For people detection, this issue can be a medium or an important one, depending on the position of all these persons. The number of persons itself in the scene is not an issue. If the scene is crowded but the persons do not occlude each others, the number of persons is not a problem. Each person is detected independently as long as their important features from the detection model are visible. On the other hand, if the important features are occluded, the detection can fail for the occluded persons.

  For the mono-camera object tracking, this issue is a critical one. The high number of persons in the scene has two negative effects: first, the probability of occlusion increases with the increase of person number. Second, the high number of persons increases the risk of permutations and tracking errors due to the high probability of having close models for tracked persons/objects.

  For people re-identification, this issue is an important one. The high number of persons in the scene increases the number of candidates for each re-identification query, and thereby, it increases the probability of error. This high number increases the probability to have similar visual signatures too.

- **Background clutter**: Sometimes background structures exhibit similar texture and shape as the one of a person or tracked objects, making distinction difficult.

  This issue does not concern directly the re-identification task as long as this last one gets its inputs (request and candidates) from the two other algorithms (people detection and mono-camera object tracking), so a failure in this task because background clutter is in reality a failure in people detection or in mono-camera people tracking.

  For people detection task, this issue is an important one. The desire to obtain a generic person model with all possible physical variations and a robust model against illumination changes is generally satisfied by ignoring some kinds of information like color or internal textures and by accepting large range of shape deformations. This enlargement of the model causes some wrong classification when a

background part has similar shape/model to the person one (a traffic signal, tree leaves and branches, etc.).

For mono-camera object tracking, this is also an issue, but less important than for people detection. There is always a risk that a tracked model clings to a background object when the tracked object passes near this background object if this last one has a high similarity with the tracked object, but thanks to the largest amount of information which can be used to model the tracked object, this risk is minimized.

- **Occlusions**: Sometimes people/objects are partially or completely occluded by objects they are carrying, by overlaps with other people/objects, or by structures in the environment.

  For people detection, this issue is important or critical depending if the occlusion is partial or full one. People detection becomes impossible by complete occlusion due to the unavailability of any information. In the partial occlusion case, this detection becomes difficult and depends on the used person model for detection. If all the features of interest of the model are visible despite the partial occlusion of the person, the detection can be performed. If only some of these features are visible, it will depends on the way the detection is performed, i.e. if an inference/extrapolation process is available. This kind of process is the most difficult to provide.

  For mono-camera object tracking, both partial an full occlusion are important issues, but the full occlusion is more important issue. In fact, in case of partial occlusion, depending on the used features and techniques for tracking, the still visible part of the tracked person/object can be sufficient to keep the tracking performing until the person/object re-appears entirely. In the case of full occlusion, this issue becomes a problem which can be solved by a collaboration between re-identification and spatio-temporal coherency. The results of full occlusion management depends on the re-identification performance, but also on many other parameters like occlusion duration, scene configuration (possible exits during occlusion), etc.

  People re-identification task is concerned by this issue only for partial occlusion case. In full occlusion situation, it means that the request/candidate person(s) are not available yet. The re-identification is performed when the request person reappears. The partial occlusion case is an important issue. If some important features of the visual signature are not visible, the re-identification fails.

- **Shadows and reflections**: Depending on illumination conditions, light angle, floor/wall smoothness, shadows and reflection can be important issues.

  For people detection and mono-camera object tracking, this issue is critical. Shadows and reflections are difficult to handle during people detection and object tracking. Depending on the features (such as motion, shape and background) used for a people detection or the object tracking, a shadow on the ground or reflected by a wall/window may behave and appear like the person/object that casts it.

  This issue does not concern directly the re-identification task as long as this last one gets its inputs (request and candidates) from both other algorithms (people detection and mono-camera object tracking), so a failure in this task due to shadow non-removal or reflection is in reality a failure in people detection or in mono-camera people tracking.

- **Different sensor response:** This issue concerns people detection and re-identification tasks.

  People detection is highly affected by this issue. State of the art people detectors, which are based on an off-line training, are strongly dependent on the condition of acquisition of training images. One of the most important acquisition condition is the sensor response. People detectors witch are trained using some specific cameras may not perform correctly in other situations where different kinds of sensors are used.

  mono-camera object tracking is a mono-camera processing. It means that the process is performed on each camera independently, and thereby, it is not affected by external (other camera) information.

  For people re-identification, this issue is a critical one. In a camera network, nothing ensures that all the cameras have the same model, and event if it is the case, the sensors may have small or large difference in their responses. The most important issue is the color response of the sensors. The same person with the same clothes can be rendered in different ways by two different cameras. Sometimes, a red pullover can be displayed as red in one camera and orange in an other one. The same for white and yellow colors. A practical example will be presented in the chapter 6.

- **Computational cost**: This is an important issue for real-time processing for live video-surveillance purpose.

  For people detection, techniques and methods that achieve state of the art detection usually require heavy computation time (for training and for detection) com-

pared to trivial person detection method. This is a difficulty especially in video surveillance context, where it is required to have reactive response acceptable for humans. Balancing detection performance with computational requirement adds to the challenges faced in person detection. Time detection can be constant or vary according to the number of persons, depending on the used technique. For example, [Dalal 2005] approach (HOG with SVM) takes the same time to test each candidate region while [Tuzel 2007] approach takes different times to test each candidate region, depending on the depth of the reached level of the cascade of classifier before rejection. Nevertheless, both kinds of approach are slow.

For mono-camera object tracking, the computational cost depends on three main parameters: first, the complexity of the used model for tracked targets. Using global color histogram is faster than computing a set of features with various information types. Second, the tracking technique. A simple Kalman filter is faster than a complex particle filter. Finally, an important parameter is the number of targets which are tracked simultaneously. The more the tracked targets are, the higher the computational cost is.

For re-identification task, computational cost is an important issue too. The complexity of the computed signature impact directly the processing time. The number of candidates for a re-identification is an other important parameter, which can make processing time explode if no filtering candidate step is performed before.

The issues mentioned above show that a successful person detection based on a single sensor is very difficult. For real world scenarios, more promising approaches combine more inputs from more than one sensory channel. In multi-modal person detection, detections from the different sensors (collaboration between visible and IR cameras, fusion of several cameras with joint fields of view, etc.) can be used to cross validate the mono-modal detection to obtain a robust detector. Features that are not captured by one sensor can be captured by another one making the detection more invariant to the above listed challenges. Unfortunately, in video surveillance context, this kind of improvement, based on multi-sensors, can not be ensured most of the time.

These issues are also significant to both single-object tracking and multi-object tracking. However, multi-object tracking also requires to solve some other issues e.g. modelling multiple object interactions. Tracking methods should be able to distinguish different objects in order to keep them consistently labelled. Although during the last few years, there has been a substantial progress towards moving object detection and tracking. But tracking an object in an unconstrained, noisy and dynamic environment still makes this problem a central focus of research interest.

A summary of the different issues and challenges, and their impact on people detection, mono-camera tracking and re-identification tasks is presented in the table 1.1

	People Detection	Mono-Camera Tracking	People Re-identification
Physical variations	$\times\times\times$		
Body/Shape deformations	$\times\times\times$	$\times$	$\times\times\times$
Illumination changes	$\times\times$	$\times\times$	$\times\times\times$
View point changes	$\times\times\times$	$\times$	$\times\times\times$
Crowded scene	$\times\times$	$\times\times\times$	$\times\times\times$
Background clutter	$\times\times\times$	$\times$	
Partial occlusions	$\times\times\times$	$\times\times$	$\times\times\times$
Full occlusions		$\times\times\times$	
Shadows and reflections	$\times\times\times$	$\times\times\times$	$\times\times\times$
Different sensor response	$\times\times\times$		$\times\times\times$
Computational cost	$\times\times$	$\times\times$	$\times\times\times$

Table 1.1: Summary of different issues and their impact on each task. $\times$: low impact. $\times\times$: medium impact. $\times\times\times$: high impact. No cross: not concerned by the issue.

1.4 Hypotheses and Constraints

This thesis has been performed with some hypothesis. Some of them are common to the three presented tasks and some others are dedicated to specific tasks. These hypothesis are:

Static cameras: Due to the industrial context of Digital Barriers, and the use of background subtraction algorithm as a provider of targets to track for mono-camera object tracking, only static cameras are considered in our study.

CIF images: For the same reason (industrial context of Digital Barriers), the necessity to run multiple analyses per server (each camera has it's own process), the network limitations (bandwidth), and for some other constraints, small size images (CIF) are used most of the time for the processing, to comply with these constraints. This implies that all the developed algorithms have to be robust enough to deal with low information amount (small people/objects sizes and resolution). When largest images are used, the performances are equivalent or better due to the availability of more information/details.

Calibrated and not calibrated cameras: All used cameras for our work are calibrated. Detected objects by background subtraction algorithm are classified using 3D

world information, and requires calibrated cameras for that. In addition, camera calibration allows to learn fast and simple real world information of tracked objects, allowing a fast occlusion management.

For people detection part, the calibration of cameras is not critical. The detailed approach can be performed on images/videos without the availability of camera calibration, but the use of camera calibration speeds up greatly the detection process in a way which is explained in the chapter 5.

Sufficient frame-rate: This constraint is concerning exclusively mono-camera object tracking task. To be able to manage all temporal variations, a minimum frame-rate of 8fps is needed.

1.5 Contributions

Our goal is to propose the innovative ameliorations on state of the art algorithms to obtain an operational (effective and efficient) framework for tracking people through a camera network. This framework has to be a turnkey system, by being as generic as possible and by do not requiring new parametrization for each deployment case while it has to provide high performances and to process in real time, due to the industrial constraints.

The presented work brings 11 significant contributions comparing to state of the art. The first contribution consists of a general framework to process the three successive steps which are ”people detection”, ”mono-camera tracking” and ”people re-identification” (See figure 1.8). One contribution is related to people detection part, four contributions to mono-camera object tracking and finally five contributions to people re-identification.

1.5.1 Contribution to People Detection

- **An optimization method to improve cascade of classifiers based people detectors:** In order to speed-up classifier training, detection task and to improve detection performances, we propose a preprocessing step which optimizes a state of the art approach. This optimization is based on clustering negative training samples before classifier training in a specific way, which can be generalized and used for all techniques which use trained cascade of classifiers. This work has been published in VISAPP 2013 [Souded 2013].

We have evaluated and compared our people detector on four datasets: INRIA, DaimlerChrysler, Caltech and CAVIAR datasets.

1.5.2 Contributions to Mono-Camera Object Tracking

- **A new method for SIFT feature detection and selection for object tracking:** We propose a new method to ensure an optimal representation of the tracked object of interest using SIFT features. This method allows better object tracking especially for partial occlusion cases. This work has been published in ICDP 2011 [Souded 2011].

- **An hybrid particle weighting method for SIFT feature particle filtering:** We use a particle filter to track all SIFT features representing the object of interest. Our proposed method allows to weight the used particles using both SIFT descriptors similarity measures and background subtraction results in a sophisticated way (not a binary weighting), dealing with various background subtraction qualities. This work has been published in ICDP 2011 [Souded 2011].

- **A data association framework for object tracking:** Once SIFT features are tracked from previous frames to the current one, our proposed data association method allows to infer object tracking state from SIFT features one, detecting and managing the several cases which may occur during object tracking (especially occlusions). This work has been published in ICDP 2011 [Souded 2011].

- **A fast occlusion management method:** We propose a fast methods to deal with full occlusion issue. it consists in learning real world information concerning the tracked object (dimensions and velocity variations), using the dominant colors extracted during the tracking and finally, using the tracked SIFT features as additional information for object re-acquisition after occlusion. This work has been published in ICDP 2011 [Souded 2011].

We have evaluated and compared our mono-camera object tracking algorithm on four datasets: PETS 2001, ETISEO, Caretaker and CAVIAR datasets.

1.5.3 Contribution to Person Re-identification

- **Fast image alignments before signature computing for multiple-shot case:** We propose a fast method to align automatically extracted images of each person before visual signature computing. The people detection and delimitation errors may provide some truncated or badly centred people images, which alter the computed signature if no processing is performed to deal with this issue. Our method allows to correct the delimitation of slightly bad delimited people, to remove the image

with significant errors, and to align the kept images to improve the computed visual signature and the comparison.

- **Use of texture information in addition to color:** The baseline method we take for our work ([Farenzena 2010]) exclusively uses color information for visual signature computation. We propose to introduce texture in addition to color in two separate ways: replacing the color characterization of the signature component (RHSP features) by a color+texture characterization using covariance descriptors, and adding SIFT features to the final signature. The SIFT features are provided by mono-camera object tracking algorithm when the whole system is used.

- **Visible side classification for more reliable signatures comparison:** People appearance may be different according to their visible side, a unique visual signature for each person may not be efficient enough. We propose a method to detect and classify the visible side of observed people in 8 classes, based on our mono-camera object tracking algorithm, and assigning a sub-signature to each observed class, and providing a more precise signature comparison method.

- **Spatio-temporal coherency filtering method:** Depending if the considered cameras share overlapping fields of view or not, we propose for each case a method to exploit the camera calibration information to reduce the number of candidates for a re-identification query by filtering incoherent spatio-temporal matching, and to weight the appearance based matching by a real world distance weight in the case of overlapping field of view.

- **Adaptive weights for signature components:** To make the re-identification approach generic, we propose an adaptive weighting method for all used features, taking in account the amount and the quality of available information (Color/Texture) and the considered people visible side if this information is available.

We have evaluated and compared our mono-camera object tracking algorithm on four datasets: VIPeR, ETHZ, iLids and CAVIAR datasets.

1.6 Outline

This PhD manuscript is organized as follows:

- **Chapter 2** presents a state of the art for each of the three studied domains: People detection, Mono-camera object tracking and People re-identification.

- **Chapter 3** presents a general overview of the whole framework for person re-identification as a complete processing chain, starting by people detection, followed by people tracking in each single camera independently, and finally people re-identification through a camera network. The dependencies and the collaboration between these three processing steps including feedback are presented. After that, an overview of the proposed approaches for each processing step is presented.

- **Chapter 4** details the proposed approach for people detection on static images and video sequences. First, a summary of Tuzel et al. [Tuzel 2007] approach for people detection and its improvement by Yao et al. [Yao 2008] are presented because our work is based on these approaches. The choice of these approaches as a basis for extension and their issues are explained and detailed, introducing and justifying our contributions, consisting in an optimization process to speed-up classifier training and detection process, in addition to improving detection performances.

- **Chapter 5** presents the proposed approach for object tracking in mono-camera context. As mentioned in the introduction, a generic object tracking algorithm is targeted first, due to the industrial needs of Digital Barriers. We highlight all specific parts restricted to people tracking when this is the case. The presented object tracking is based on SIFT features using Particle Filtering approach, followed by a data association reasoning stage to lead to a complete object tracking. Several contributions are presented at different levels of this process.

- **Chapter 6** presents the proposed improvements to state of the art algorithms for people re-identification. The baseline approach is explained, and its limitations are highlighted. The approach is based on the symmetry-driven accumulation of local features. The contributions are detailed and tested to show the improvements they provide.

- **Chapter 7** is dedicated to experimental results and benchmarking with state of the art studies. Each of the presented processing steps is evaluated on several dedicated benchmarking datasets and compared with state of the art approaches. The proposed approaches and contributions of each processing step are validated by comparing with the state of the art approaches performances.

- **Chapter 8** presents the concluding remarks and limitations of the thesis contributions for each processing step. We also discuss about the short-term and long-term perspectives of this study.

2

STATE OF THE ART

This chapter presents the state of the art of the three topics covered in this thesis.

The first section concerns the state of the art in people detection on images and video sequences. This separation between static images and video sequences is mainly related to the type of used features due to the additional information provided by the movement. The state of the art in people detection can be divided into three categories, depending on the processing step for detection: candidate region selection, pertinent feature extraction and learning/classification techniques.

The second section concerns the state of the art in mono-camera object tracking, which can be classified according to two criteria: ”how to represent/model the tracked object” (i.e. which features to characterize it) and ”how to update search for its model over time?”. The object modelling is performed in several ways, according to the used information: color, shape, texture and motion. Tracking techniques can be divided into two main categories: deterministic and probabilistic approaches.

The last section concerns the state of the art in people re-identification. Approaches are divided into two main families: biometric and appearance-based approaches. The biometric approaches include iris recognition, fingerprint recognition, face recognition and gait recognition. The appearance based approaches are classified into single-shot approaches and multiple-shot approaches, according to the number of used images per person. In both of them, the appearance modelling is performed following two main kinds of methods: Feature oriented methods and Learning methods. Some recent approaches use context information instead of focusing on people on interest images only.

2.1 People detection

People detection on static images and video sequences is a critical task in many computer vision applications, including human-robot interactions, robot navigation in presence of humans, pedestrian detection for Automated Driver Assistance Systems (Stereo vision-based pedestrian detection system is included in both the new 2013 Mercedes-Benz E-Class and S-Class models, since june 2013), content based image and video processing, and in a large proportion, video surveillance which is our topic of interest.

At the same time, it is one of the most challenging problems in computer vision due to the large number of possible situations, including variations in people appearance and poses. These challenges are further augmented in video surveillance applications due to computational time requirements for a reactive system, especially when this detection is not the final aim, but one step in a more complex processing like in our case.

Due to its importance and its several challenges, People detection is an active research area with a rapid rate of innovations.

In this section, different people detection approaches from the state of the art are presented. People detection methods can be categorized into two main families: Trained classifier approaches and template matching approaches.

The trained classifier based approaches are performed in two distinguished steps: training and detection. The training step focuses on the significant features extraction using several machine learning methods to obtain a person-class model. The obtained model is then used for detection using different ways.

The template matching approaches aims at extracting a generic template for person's class using pertinent features. Detection is performed using a direct template matching procedure, using several proposed matching measures and distances.

To help structure the flow of this chapter, figure 2.1 shows a typical flow diagram of training based people detectors procedures. In the training step, discriminative and pertinent features are extracted from a training dataset and labelled according to their class (positive or negative data). A classifier is then trained on these data using several machine learning techniques, providing a person class model. For detection step, images for the observed scene are acquired, then the searching areas on the images are defined as the candidate regions, either by searching on the whole image with any a priori knowledge, or by targeting specific regions using real world information. From these candidate regions, discriminative and pertinent features from the same type than the used ones for training are extracted for classification. The trained classier is finally used to decide whether the evaluated candidate region corresponds to a person or not.

In the following paragraphs, we present the used techniques in the main state of

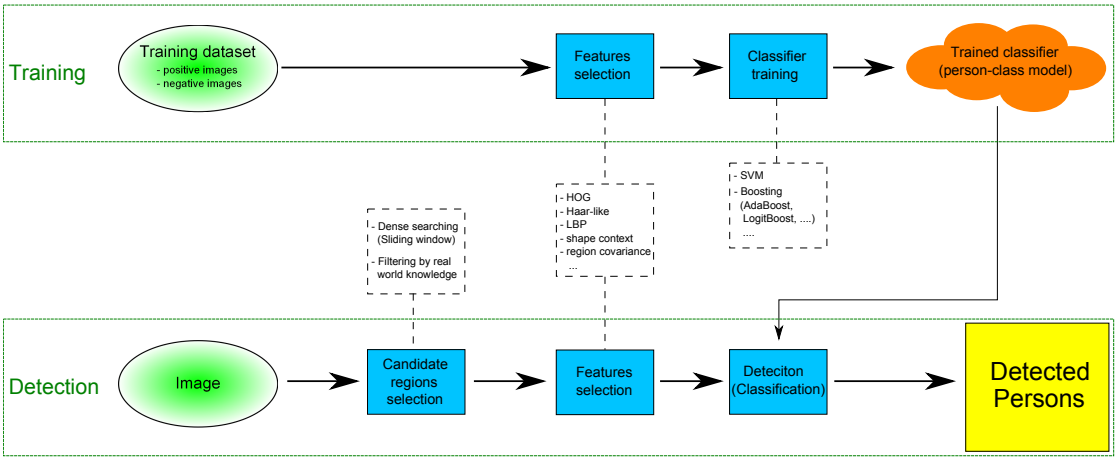


Figure 2.1: Successive steps for people detector training and detection.

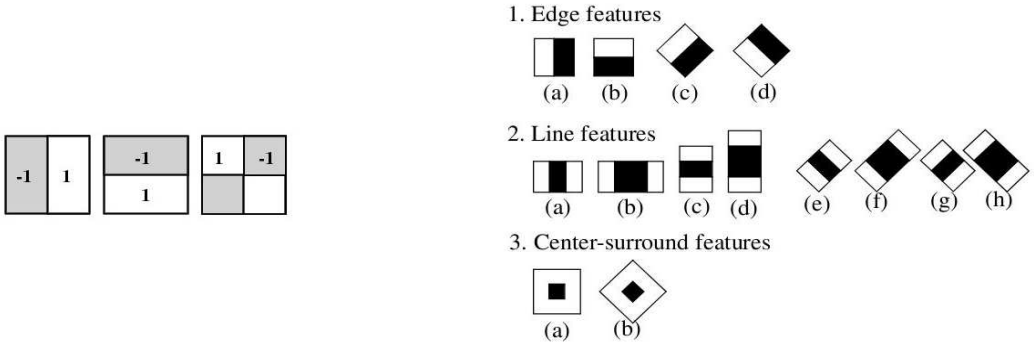
the art approaches, starting by exposing the most used features for person class modelling, the two candidate region selection methods and finishing by the classification techniques.

2.1.1 Pertinent Features

An image consists in a array of pixels. Using each pixel independently does not allow to extract any information about image content. For this reason, many approaches for visual features extraction from groups of pixels are proposed in the state of the art. These features enable us to extract meaningful information from image. In this section, different image features that are used in visual person detection are briefly presented.

2.1.1.1 Haar-Like Features

Haar Wavelets were first introduced in the context of Object Detection in late 90s by Papageorgiou et al. [Papageorgiou 1998]. Viola and Jones [Viola 2001] adapted the idea of using Haar wavelets and developed the so-called Haar-like features. They introduced the notion of Integral Image so as to compute these features in a fast way. The Haar-like features encode the relationships between average intensities of neighbouring regions along different orientations capturing edges or changes in texture. This makes them suitable to capture the structural similarities between various instances of a class. Figure 2.2a shows the three types of 2-dimensional Haar-like features used by [Oren 1997]. These features capture change in local intensity along horizontal, vertical and diagonal directions. When applied to images, the value of a two rectangle feature is the difference between the sum of the pixels lying in the unshaded area with the sum



(a) Three of the Haar-like features used by [Oren et al., 1997], (horizontal, vertical, and diagonal)

(b) Complete set of extended Haar-like, including center surround, feature templates used by [Lienhart et al., 2002]

Figure 2.2: Haar like features templates.

of pixels lying in the shaded area. A four rectangle feature computes the difference between diagonal pairs of rectangles.

Leinhart et al. [Lienhart 2002] introduced a set of extended haar-like features by adding upright, 45 °C oriented and center-surround rectangular features allowing the prototypes to be scaled independently in vertical and horizontal axis. Figure 2.2(b) shows the complete Haar-like feature template used by [Lienhart 2002].

2.1.1.2 Edge Orientation Histograms (EOH)

Silhouette and edge information are important features to discriminate a person in images. To encode these information, Edge Oriented Histograms (see figure 2.3) have been proposed initially for face detection by Levi and Weiss [Levi 2004]. These features not only maintain invariance to global illumination changes, but also capture geometric properties that are difficult to capture with other features. Later, Edge Oriented Histograms have been used for people detection. In [Gerónimo 2007], a combination between Haar-like features and Edge Oriented Histograms is used as discriminant feature for classification.

2.1.1.3 Histogram of Oriented Gradients (HOG)

Another feature for silhouette and edge information encoding, called Histogram of Orientation Gradients, is proposed by Dalal and Triggs in [Dalal 2005] for people detection. The feature extraction is more complex than in Edge Orientation Histograms, increasing the discriminative power of the descriptor while ensuring a certain degree of invariance. As described in [Dalal 2005], HOG descriptor computation is done in five

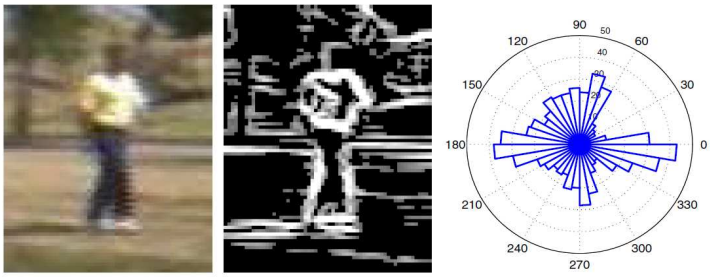


Figure 2.3: Edge orientation histogram. (Left) Example image. (Center) Edge strength image. (Right) Polar plot of edge orientation histogram. (souce [Yang 2005]).

steps:

1. A global image normalization equalization, using a gamma compression, is performed to reduce the effects of local shadowing and influence of illumination variation effects.

2. Computation of first order image gradients.

3. The image window is divided into small spatial regions, called “cells”, and a local 1D histogram of edge orientations with K orientation bins over all the pixels in the cell is accumulated. Each edge pixel contributes to each orientation bin with a value proportional to the magnitude of its orientation.

4. A normalization step is carried out by accumulating a measure of local histogram “energy” over local groups of cells called “blocks”. Each cell is normalized with respect to the block which it belongs.

5. The final HOG descriptor of the whole detection window is obtained by concatenating all HOG descriptors of all blocks of a dense overlapping grid.

The HOG feature extraction is depicted in Figure 2.4 taken from [Dalal 2005].

Four variants of the HOG descriptor have been presented by the authors. The difference between them lies in the shape of considered cells. These four variants are: Rectangular HOG(R-HOG), which is the original one, Circular HOG(C-HOG) where the cells are defined into grids of log-polar shape. Bar HOG where the descriptors are computed similar to the R-HOG, but use oriented second derivative filters rather than first derivatives and Center-Surround HOG which use a centre-surround style cell normalization scheme.

Many other approaches for people detection, using HOG descriptors, have been proposed. We can cite [Zhu 2006b, Corvee 2010, Bertozzi 2007]. They mainly differs in the

way that the HOG descriptor is used (describing the whole person image or body parts independently) or in the used classification method (SVM, boosting, etc.).

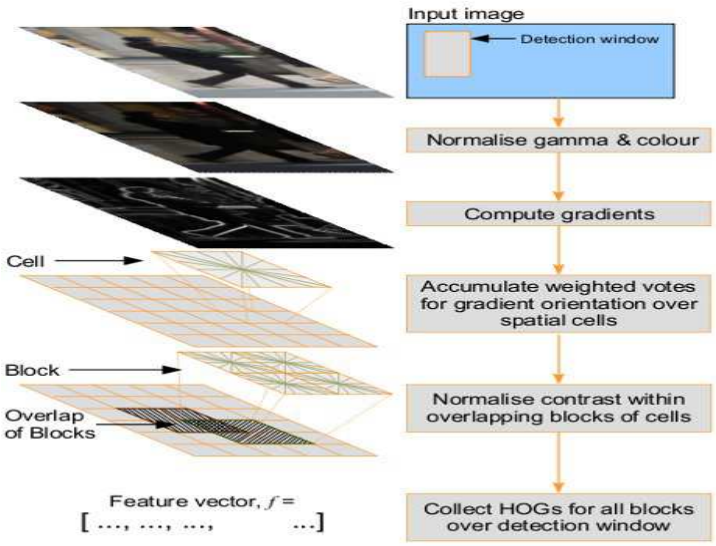


Figure 2.4: HOG feature extraction steps (From [Dalal 2005])

2.1.1.4 Local Binary Pattern (LBP)

Local Binary Patterns is texture encoding feature. It is a particular case of the Texture Spectrum model proposed in [Wang 1990] and [He 1990]. It has been first described in [Ojala 1996]. The original version of the local binary pattern feature for each pixel is based on a 3×3 pixel block of an image. The pixels in this block are thresholded by its center pixel value, multiplied by powers of two and then summed to obtain a label for the center pixel. As the neighbourhood consists of 8 pixels, a total of $2^8 = 256$ different labels can be obtained depending on the relative gray values of the center and the pixels in the neighborhood. A more generic LBP feature is proposed in [Ojala 2002]. It allows more information extraction from variable circular neighbourhood of the center pixel, according to two parameter which are the radius of circular neighborhood “R” and the number of considered neighborhood points “P” (see figure 2.5).

Many people detection approaches are based on the use of LBP descriptors only, or in collaboration with other features, improving the classification performance rates. In [Mu 2008], Semantic LBP (S-LBP) and Fourier LBP (F-LBP), two new variants of LBP feature, are proposed and used for human detection. Semantic LBP (S-LBP) feature is computed by binarizing the image on a color space such as CIE-LAB. Neighbors whose

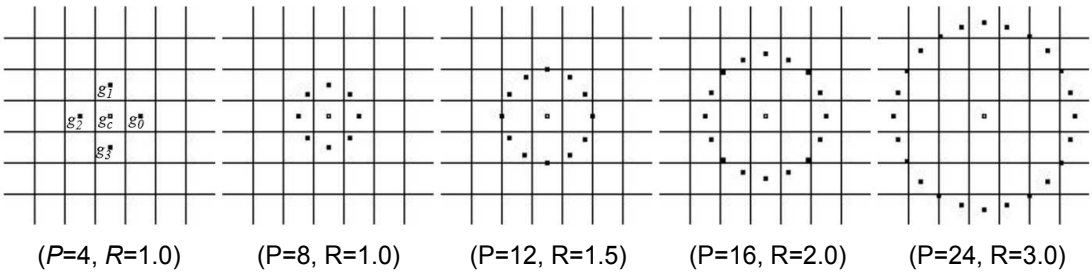


Figure 2.5: Generic LBP: Examples of circular neighbour sets for different *(P, R)* ([Ojala 2002])

distances to the central pixel exceed local threshold are marked as "1", else "0". Then arches number is counted and non-uniform arches (i.e. having more than one arches) are abandoned. The 2D histogram descriptor for any image region can be obtained by collecting information from all its inner pixels. The final feature vector is obtained by concatenating each column of the 2D histogram to get a 1D vector (see figure 2.6). The Fourier LBP is designed via similar idea ot Fourier boundary descriptor [Gonzalez 2001]. First, color distances between the considered neighbourhood pixels and the center one are grouped in a raw feature vector. Then, this raw feature vector is transformed into frequency domain. Coefficient for low frequencies are kept and used for F-LBP representation since they capture salient local structures around the center pixel.

in [Zhou 2012], standard HOG and LBP features are extracted from Regions of Interest (ROI) of human body, and are combined to characterise persons, providing better detection performance in comparison of using each descriptor independently. The classifier is trained with simple linear SVM.

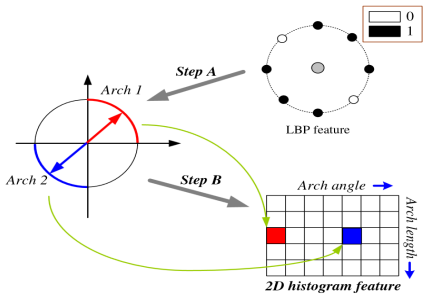


Figure 2.6: [Mu 2008] S-LBP computing method. Step A: Calculate principle directions and lengths for each arch. Step B: Vote for corresponding histogram bins.

2.1.1.5 Shape Context

Shape Contexts were first introduced by [Belongie 2002] in the context of object recognition. The approach consists in picking n points on the contours of a shape, extracted using an edge detector. Then the edges are stored in the bins of a log polar histogram formed by quantizing the locations around the picked points in both radial and angular directions. Orientation is then quantized in pre-defined number of bins. By making the location bins uniform in log-polar space, the descriptor can be made sensitive to nearby sample points more than those points further away. These descriptors are very well suited for matching purposes and have also been used for pedestrian detection by [Leibe 2005].

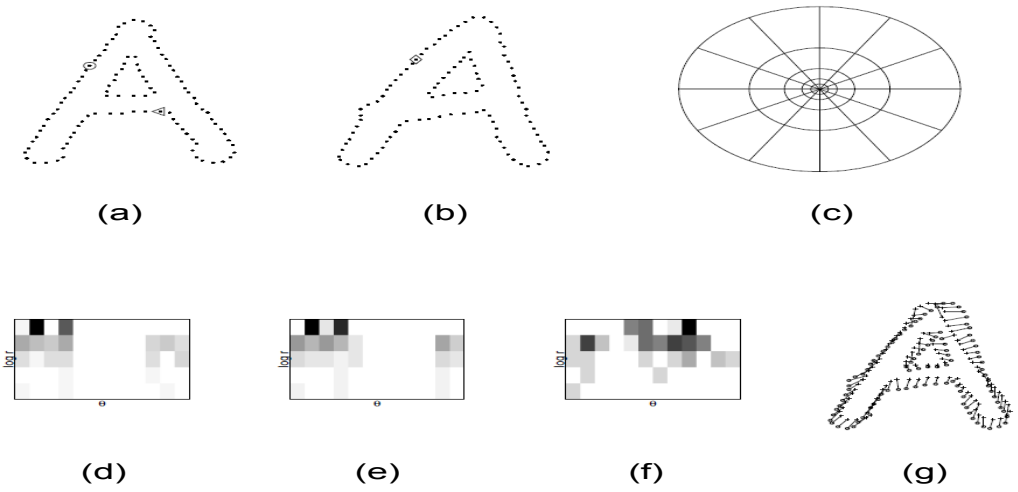


Figure 2.7: [Belongie 2002] Shape Context computation and matching. (a,b) Sample edges points of two shapes. (c) Diagram of log-polar histogram bins used in computing the shape context. 5 bins for $\log r$ and 12 bins for θ are used. (d,e,f) Example shape contexts for reference samples marked by $\circ, \diamond, \triangleleft$ in (a,b). Each shape context is a log-polar histogram of the coordinates of the rest of the point set measured using the reference point as the origin. (Dark=large value.) Note the visual similarity of the shape contexts for $\circ, \diamond$, which were computed for relatively similar points on the two shapes. By contrast, the shape context for $\triangleleft$ is quite different. (g) Correspondences found using bipartite matching, with costs defined by the χ^2 distance between histograms.

2.1.1.6 Region Covariance Descriptor

Region covariance descriptors have been first introduced for people detection by Tuzel et al. in [Tuzel 2006]. Each image pixel can provide a large amount of basic infor-

mation like its coordinate, its intensity and color values, its first and second derivative values according to X and Y axis and many other values obtained by applying several filters and operators. By extracting a d-dimensional vector of features from each pixel of a given image region, it is possible to compute the variance of each type of considered feature and the correlation between these different features, and encapsulate all these information in a single d-dimension square matrix which will describe the considered region. More details concerning the way to compute this descriptor and its characteristics, in addition to all the related metrics are presented in the chapter 4 to avoid information redundancy, since our proposed method is based on this descriptor.

In [Tuzel 2007], 8-dimensional covariance descriptors are used for the people detection training. The computation of these region covariance descriptors is speeded up using Integral Images. A cascade of classifiers is trained using a boosting scheme in Riemannian Manifold, due to the nature of covariance matrices. This approach provide interesting detection performances but it requires non-negligible processing time, due essentially to the eigenvectors decomposition which the basis of all metrics computation, and which is directly proportional to the region covariance descriptors dimension.

Yao and Odobez improve [Tuzel 2007] method by introducing three modifications to the initial approach in [Yao 2008]. First, the second derivative features on X and Y axis which were used in [Tuzel 2007] 8-dimensional features are replaced by two other features, related to the background subtraction, speeding up the detection in video sequences. Then, they greatly speed up the detection by building classifiers with 4-dimensional region covariance descriptors instead of the 8-dimensional covariance descriptor in [Tuzel 2007], using the best subset of 4 features from the initial set of 8 features, for each considered region of interest. Finally, they increase the discriminative power of extracted features for each region of interest by concatenating the mean feature vector of the same region upon all positive training images to the projected covariance matrices on vector space before regression computation. These improvements, their justification and the way in which they are performed are detailed in chapter 4.

2.1.2 Candidate Region Selection

This processing step concerns the detection process only as long as the positive training dataset are provided by labelled person images.

The candidate region selection is the first step for person detection in images. In general, images contain various environment objects. They may contain none or many people. In the case of images containing people, a person is defined by a sub-image region containing a set of specific features. It is therefore necessary to select and test “some” candidate image regions to check if they correspond to persons or not. This can-

didate selection presents two main advantages: First, the processing time would be decreased by checking only selected regions and second, possible false positives would be filtered out by this selection. In the same time, this selection can lead to miss-detections if relevant image regions are not checked.

Candidate region selection is done according to specific criteria and can be divided into two categories: Dense searching by sliding window and real world a priori knowledge (camera calibration).

2.1.2.1 Dense Searching by Sliding Window

This method is generally used when no information is available concerning the scene in the image. It consists in scanning the whole image using a sliding window (see figure 2.8 (a,b)). The size of the scanning window as well as the displacement step are dependent on some criteria like the sensitivity of the used features with respect to small/large shifts and rescales. To ensure the detection of people with different sizes in the image, a multi-scale scan is performed. Some approaches rescale the image by keeping the search window with a constant size, corresponding to the positive training images size, while others approaches rescale the scanning windows to detect people at various scales on the same image.

The choice between these two policies depends on the constraints on the types of used features and classifiers. For example, Dalal and Triggs [Dalal 2005] construct an image pyramid by scaling the input image by a factor of 1.2 and use a scanning window with a constant size of 64x128 pixels. They shift this scanning window by 8 pixel in both axis (the scanning shift is constrained by HOG cell dimension which is 8 pixels. Only multiples of cell dimension can be used as browsing step). The size of scanning window corresponds to the size of the images which were used to train a SVM classifier. Due to the nature of the SVM classifier, it is not possible de rescale the scanning window because the corresponding SVM classifier has to be adapted to the new window size, which is not a trivial operation (SVM hyperplane dimension is dictated by the dimension of the whole image HOG feature vector, which depends on the image size).

In opposition to this example, Tuzel et al. [Tuzel 2007] have tested and compared both possibilities in their approach: First they rescaled input image at different scales and applied a scanning window with constant size on each scaled image, and second, they keep the original image for scanning and applied different detection window sizes. In their approach, the nature of region covariance descriptor that they used as a feature allows to rescale classifier with a negligible variation. This is due to the fact that the information contained in a given region in term of feature variances and correlation between them does not vary by a uniform rescaling on both axis theoretically. In practice,

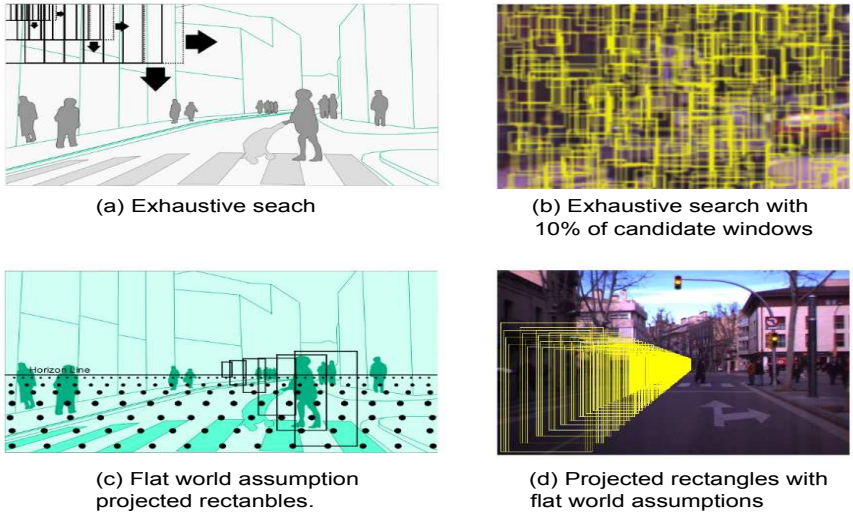


Figure 2.8: Illustration of sliding window and flat world assumption for candidate regions selection ([Gerónimo 2009]).

a negligible variation occurs du to the discrete (non-continuous) nature of image pixels, and the information modification introduced by anti-aliasing filtering after rescales (bilinear filtering for example).

Note that it is better, when it is possible, to rescale the scanning window (and adapt the corresponding classifier) instead of rescaling input image. The processing time can be highly reduced by this choice. In fact, rescaling input images, in addition to the necessity to re-extract the whole image raw feature vector at each scale, are high time consuming if the number of scales is important. This rescaling time must be multiplied by the number of images in case of videos.

Rescaling scanning window is faster, even by training a classifier on one images size only and rescale this classifier directly using its mathematical properties (e.g. region covariance descriptors) or by training several classifiers on predefined set of image sizes (predefined set of detection scales). Even if the classifier adaptation (rescaling) requires computational time, this operation can be done off-line, before the on-line processing. Unlike input images content that change over the time (and thereby the contained feature values), a classifier for a given scanning window size is constant (it corresponds to a fixed model). All the needed classifier scales are computed and stored before processing and directly used during detection process.

2.1.2.2 Filtering by Real World Knowledge

This approach uses a kind of scanning window similar to the previous method, but the main difference is that scanning is not performed on the whole image, and the searching scales are highly reduced.

This approach is strongly constrained by the camera calibration information availability, and is based on the assumption that people are on the ground floor. For a calibrated camera, rectangular regions corresponding to the aspect ratio of a person are placed on the ground floor of the 3D world up front and projected onto the image using camera transformation matrix. These regions then constitute the candidate windows for further processing (see figure 2.8 (c,d)). These approach has been applied in pedestrian detection from a vehicle by Gavrila et al. [Gavrila 2004] and Gerónimo et al. [Gerónimo 2006]. Using this approach, the number of possible regions for subsequent steps is highly reduced. Gerónimo et al. [Gerónimo 2006] have shown that the performance of this candidate generation scheme is very dependant on the accuracy of the camera calibration parameters.

2.1.3 Classification

The classification is the final step of people detection. During the classification, a candidate region is evaluated and a decision is taken whether it is a person or not. State of the art of people detection is dominated with classifier training approaches. Most of them use variants of Boosting and SVMs machine learning. But a silhouette matching technique known as Chamfer System has also been used.

Both of these kind of classification methods are presented in the following paragraphs.

2.1.3.1 Chamfer Matching

Chamfer Matching, introduced by [Barrow 1977], is a technique used to compare the shapes of two collections of shape fragments. For example, for an edge template T composed of edge features t and an image's edge map I, the Chamfer Distance is given by the average distance dI to the nearest feature.

$$D_{Chamfer}(T, I) = \frac{1}{|T|} \sum_{t \in T} d_I(t) \tag{2.1}$$

Depending on how well the template represents the person's class and how good the features used are, this measure can be used for people detection. In order to make a precise decision about the object location, orientation, and scale, it may be necessary to use

subsequent verification stage [Gavrila 2004]. This method has been applied successfully for person detection from images [Gavrila 2004].

2.1.3.2 Support Vector Machines (SVMs)

Support Vector Machines are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis. they were introduced by Vladmir Vapnik [Vapnik 1995]. SVMs are widely used in people detection approaches due to their high generalization performance without any needed a priori knowledge, even when the dimension of the input space is very high [Vapnik 1995].

The aim of SVM machine learning is to find the optimal hyperplane that separates two classes of data. Depending on the nature of data, this separation may be a linear or non-linear one. The most used non-linear kernels for SVM are polynomial and Gaussian kernels, but most of approaches for people detection assume that the class separation can be performed well with linear SVM's.

In the case of linear SVM, considering a set of N linearly separable training examples $(x_i, y_i)_{i:1..N}$, where $x_i \in R^p$ are p-dimensional real vectors and $y_i \in \{-1, 1\}$ their corresponding class labels, an infinite number of separation hyperplanes can be taken (see figure 2.9(a)). SVMs aim to select the optimal hyperplane for separation. It is done by introducing "maximum-margin" notion.

Any hyperplane can be written as the set of points $\mathbf{x}$ satisfying $\mathbf{w} \cdot \mathbf{x} - b = 0$ where "$\cdot$" denotes the dot product and $\mathbf{w}$ the normal vector to the hyperplane. The parameter $\frac{b}{\|\mathbf{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$.

The margin maximization is performed by finding the furthest two parallel hyperplanes which delimits a points-free space between them (see figure 2.9(b)). These two hyperplane are defined by the two equations $\mathbf{w} \cdot \mathbf{x} - b = 1$ and $\mathbf{w} \cdot \mathbf{x} - b = -1$. The distance between these two hyperplanes is $\frac{2}{\|\mathbf{w}\|}$, so maximizing this distance is done by minimizing $\|\mathbf{w}\|$. To prevent data points from falling into the margin, this minimization have to be done under the two following constrains:

$\mathbf{w} \cdot \mathbf{x}_i - b \geq 1$ for $\mathbf{x}_i$ of the first class, and

$\mathbf{w} \cdot \mathbf{x}_i - b \leq -1$ for $\mathbf{x}_i$ of the second class.

These two constraints can be grouped, using the class labelling, in the following equation:

$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1, \qquad \forall 1 \leq i \leq N$

The optimal separation hyperplane is taken as the one between the two parallel hyperplanes, at equal distance from each of them.

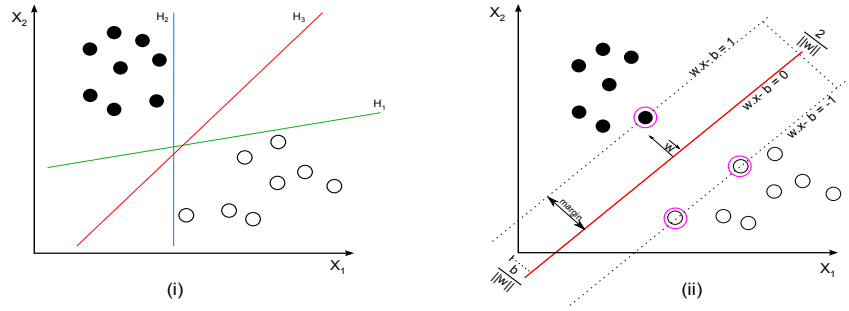


Figure 2.9: Linear SVM illustration with 2-dimensional points separation. (i) H_1, H_2, H_3 are three sample plans from the infinite possible plans which can separate the two point classes. (ii) Optimal plan(in red) is taken, by maximizing the margin between Support Vector points (surrounded by purple circles).

Note that a variant of SVMs, called Relevant Vector Machines (RVM) has been proposed in [Tipping 2000] and [Tipping 2001]. The RVM has an identical functional form to the support vector machine, but provides probabilistic classification.

Several approaches for people detection, based on SVM training were proposed. In [Dalal 2005], people detection is performed by a linear SVM classifier trained on Histogram of Oriented Gradients features. In [Miezianko 2008], a linear SVM classifier is trained on gradient patches extracted from low resolution infrared videos. In [Papageorgiou 2000], haar wavelets are used to train a polynomial SVM.

In [Ronfard 2002], a body part based people detector is trained using Relevant Vector Machines (RVM). It will be presented in subsection 2.1.4.2.

2.1.3.3 Boosting

Boosting is a field of machine learning domain. The principle consists in iterative addition of trained weak classifiers to perform a strong classifier. A weak classifier is defined as classifier which can separate two classes at least as well as a random classifier, i.e. it does not exceed 50% of an average error rate if the classes are equivalently distributed. Each added weak classifier is weighted in the strong classifier according to its classification quality: the more it classes well, the more important is (higher his weight is).

The name of this machine learning method comes from its main common processing step: training samples witch are incorrectly classified after a given iteration are "boosted" by assigning them more important weights in the next iteration step, to make the machine learning "focus" more on them.

The final strong classifier consists in a set of weighted weak classifiers, and its clas-

sification decision is performed using the sum of weighted responses from all its weak classifiers.

Many variants of Boosting algorithms were proposed, we can cite Bootstraping, AdaBoost, Discrete AdaBoost, Real AdaBoost, GentleBoost, BrownBoost, LogitBoost, AsymBoost, KLBoost, FloatBoost, GloBoost, RankBoost, etc.

As the subject of our thesis does not relies to machine learning technique, we can not detail all these algorithms. We can only indicate that they all share the basic principle, but differs on the way they perform some computing step, like the new weight update for incorrectly classified training samples, the weak classifier selection criteria, the type of data for which they are dedicated and the number of initialization parameters that they requires.

Nevertheless, and due to its large use in many people detection approaches, we will focus on AdaBoost algorithm and its very close variants LogitBoost and RealAdaBoost.

AdaBoost (Adaptive Boosting) has been introduced by Freund and Schapire in [Freund 1995]. The following paragraph summaries it's algorithm.

For a binary classification, let us consider a set of N training p-dimensional samples $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in R^p$, $y_i \in Y = \{-1, +1\}$. The number of weak classifiers to train, which correspond the number of iterations of the main algorithm's loop, is fixed at the beginning. we will note it T.

- The first step consists in initializing the weight distribution D_i of all training samples: $D_1(i) = \frac{1}{N}$,.
- For each iteration t = 1,...,T:
  - A set $\mathcal{H}_t$ of q weak classifiers $h_t^{(j)}$ is formed first (j=1,...,q). The corresponding weighting error $\epsilon_t^{(j)}$ of each weak classifier $h_t^{(j)}$ is computed with respect to D_t by:
    $\epsilon_t^{(j)} = \sum_{i=1}^{N} D_t(i) I(y_i \neq h_t^{(j)}(x_i))$, while I is the indicator function.
    The best weak classifier h_t is then selected as the one with the maximum absolute value of the difference of the corresponding weighted error rate ϵ_t and 0.5: $h_t = \underset{h_t^{(j)} \in \mathcal{H}_t}{\operatorname{argmax}} \left| 0.5 - \epsilon_t^{(j)} \right|$
  - If $|0.5 - \epsilon_t| \leq \beta$, where β is a previously chosen threshold, then stop.
  - Choose $\alpha_t \in \mathbb{R}$, typically $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$.
  - Update all sample weights. For i = 1,...,N:
    $D_{t+1}(i) = \frac{D_t(i) \exp(\alpha_t I(y_i \neq h_t(x_i)))}{\sum_i D_t(i) \exp(\alpha_t I(y_i \neq h_t(x_i)))}$
    where the denominator is the normalization factor ensuring that D_{t+1} will be a probability distribution.

The final binary classifier is then obtained as: $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

AdaBoost has proven to be very useful in feature selection. At the end of the training, the features with prominent discriminative power have higher weights. It is frequently used for image classification as a result of its simplistic implementation, very good feature selection abilities, and fairly good generalization.

It has been used for person detector in many approaches. Note that all proposed approaches train final classifiers which can be either a single strong classifiers or a cascades of strong classifiers.

Cascade of classifiers perform in a rejection scheme. It is implicitly associated to a class of interest (the positive one). The tested data is evaluated by the cascade levels in increasing order. If any cascade level returns a negative response, the data is immediately considered as not belonging to the class of interest and is rejected. Is is not evaluated by the remaining levels. Only positive data (belonging to the associated class) fully pass through the whole levels of cascade when they are evaluated.

In both [Zhu 2006b] and [Jia 2007], a cascade of classifiers is trained using AdaBoost algorithm on Histogram of Oriented Gradients.

Other Adaboost derivatives method are used for the same purpose. In [Gerónimo 2007], a Real AdaBoost is used train a classifier based on a combination between Haar-like features and Edge Oriented Histograms. Chen et al. [Chen 2007b] people consists in a cascade of Real AdaBoost trained classifiers, using Edge Oriented Histograms features.

LogitBoost is used by Tuzel et al. [Tuzel 2007] and by Yao et al. [Yao 2008] to train a cascade of classifiers based on Region Covariance Descriptors. Note that in these approaches, the LogitBoost algorithm has been slightly modified to deal with the Riemannian manifold which Region Covariance Descriptors belongs.

Real Adaboost differs from the original Adaboost in the way than the original Adaboost classifier returns only a binary value (response) for a tested data, depending on this data belongs to the the first or to the second class. The Real Adaboost algorithm provide a real valued probability of class membership.

LogitBoost algorithm is first presented in [Friedman 1998]. The authors have introduced more direct approximations to AdaBoost algorithm and have shown that they exhibit nearly identical results to boosting while the computation cab be reduced by factors of 10 to 50. The main difference between Adaboost an Logitboost resides in the way the weak classifier errors are computed and thereby the way the best weak classifier is selected at each iteration. AdaBoost minimizes an exponential loss function while LogitBoost minimizes a logistic loss (hence its name).

2.1.4 Person Detection Approaches

In the previous sections, we have presented the several used features and the different classification/matching methods. People detectors can also be categorized into two major types, one which searches the image for full human bodies by scanning the input images in different ways, known as full body detection approaches, and the other which tries to aggregate evidence of existence of a person by modelling and detecting human body models, known as part based approaches. Some approaches are designed to detect the full body as well as the body-parts([Felzenszwalb 2010, Wu 2005])

2.1.4.1 Full Body Detection

In full body detection, the input image is scanned for people searching using a window with an average human aspect ratio. Most of the works in full body detection were done either from a pedestrian detection context [Gavrila 2004, Gerónimo 2006] or from a general object detection framework context [Dalal 2005].

In pedestrian detection, the fist promising results were reported by Oren et al. [Oren 1997]. In this approach, Haar-like templates are used to extract features, an SVM classifier is trained on a selected subset of them. Only significant features for the task, identified using a template learning stage, are used for training.

Jones et al. in [Jones 2003] improve [Oren 1997] method by incorporating motion information to detect pedestrians. Feature vectors consist in motion information as well as intensity information, extracted using Haar-like features. AdaBoost is used to train a cascade of classifiers. This approach outperforms previous one, but due to the use of motion information, this approach can be only applicable to static camera.

In [Gerónimo 2006], different extended Haar-like filter sets are combined with Edge Orientation Histograms (EOH) to model human body. Real AdaBoost is used for classifier training. The person detection is optimized by restricting the search area at specific image locations, determined by estimating the current ground plane. Once the ground plane is estimated, a 3D grid, sampling the road plane is projected on the 2D image defining candidate regions (subsection 2.1.2.2). The results showed that the proposed approach improves detection performances, due to two reasons. First the combination of extended Haar-like features with the Edge Oriented Histograms provides better results thant the use of Haar-like features only. Second, the ground plane estimation improves the true pedestrian location hypothesis.

In a similar context, Monteiro et al. [Monteiro 2007] used Haar like features to detect pedestrians with an AdaBoost trained cascade of classifier. In this approach, pedestrian detection is done by a sliding window approach, and the multi-scale search is

performed by and scaling the detector rather than the image.

Other approaches for person detection are performed in a general object detection context. In [Dalal 2005], HOGs as features and a Linear SVM as a classifier, and a sliding window approach for candidate generation are used. The major contribution was the construction of particular effective features, HOGs, and a data-mining approach during training in which resulting false positives were re-introduced as hard negative examples. This approach was the winner of the 2006 PASCAL object detection challenge [Everingham 2009].

Zhu et al. [Zhu 2006b] reformulated [Dalal 2005] detector in cascade of classifiers method to achieve fast and accurate human detection system. In this approach, variable block sizes HOG features are used. The selection of the important sets of blocks is performed using AdaBoost. Their final implementation used an integral array representation to speed-up the detection in comparison to [Dalal 2005] detector while maintaining similar performance levels.

Laptev [Laptev 2006] used the AdaBoost framework to select prominent features for object detection. Weighted local histogram of gradient orientations in all rectangular sub-window of the object are used as features. Weighted fisher discriminant analysis is used as a weak classifier, and AdaBoost selects the best features out of all histograms computed on all sub-windows.

2.1.4.2 Body Parts Based Detection Approaches

Body parts based detection approaches are performed in two steps. First, the different parts of the body (head, face, arms, torso, leg) are detected. Then, an inference process using the detected parts and/or some geometrical constraints is performed to deduct the presence and position of a person. Performing these two steps require to model human body parts using visual features and classifiers, and to model the topology of these body parts.

In [Forsyth 1997] body plans for people and animals detection in images are introduced. Body plans model people as an assembly of cylindrical parts, each cylinder corresponding to part of a body, where the individual geometry of the parts and the relationship between parts are constrained by the geometry of the skeleton and ligaments. A human Body Plan is constructed by segmenting human skin using color and texture criteria, assembling the extended segments, and using a hand built body plan to support geometric reasoning. In [Mikolajczyk 2004], the authors have reported that this body detector fails in the presence of clutter and loose clothing.

In [Felzenszwalb 2000], authors propose a parts based detector using pictorial structures. A pictorial structure is a collection of parts arranged in a deformable configura-

tion. Each deformable configuration is represented by spring-like connexions between pairs of parts. The parts are modelled as rectangle with fixed aspect ratio, average color, and color variance. The authors propose an algorithm for finding the global match of pictorial structure to an image.

Due to the use of simple color as feature for parts characterization, the proposed detector in [Felzenszwalb 2000] is not really robust. Ronford et al. [Ronfard 2002] improve the previous method by using better part features and detectors. A hand-labelled articulated body model with 14 joints and 15 body parts and a feature set consisting of Gaussian filtered image and its first and second derivatives are used. 15 body-parts detectors are trained using Support Vector (SVM) Machines and Relevance Vector Machines (RVM), and are applied for people detection (see figure 2.10).

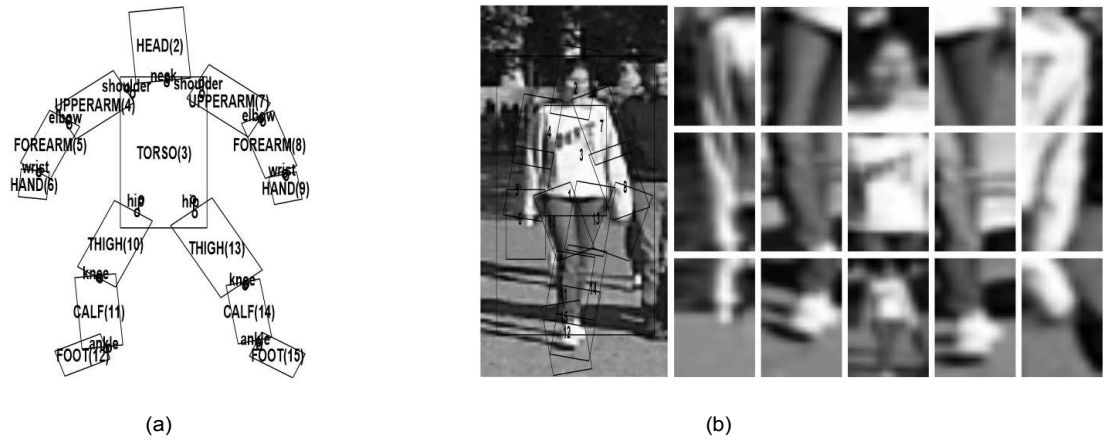


Figure 2.10: [Felzenszwalb 2000] body parts based people detector. (a) Articulated body model with its 14 joints and 15 body parts (the whole body is considered as a part, but not displayed in the figure for illustration clarity). (b) A hand-labelled training image and its extracted body part sub-images. Reading vertically from left to right: left upper arm, forearm, hand; left thigh, calf and foot;head, torso and whole body; right thigh, calf, foot; right upper arm, forearm and hand

In [Mikolajczyk 2004], Humans are modelled as flexible assemblies of parts. These parts are represented by co-occurrences of local features which capture the spatial layout of the part's appearance. Features selection and the part detectors are learnt from training images using AdaBoost. In total 7 different body parts (frontal head, frontal face, profile head, profile face, frontal upper body, profile upper body, and legs) are used (see figure 2.11). The geometric relationship between body parts is represented by a Gaussian and its parameters are learned from the training set. Dominant orientations based on first and second derivatives over a neighbourhood computed at different scales and combined, each three neighbouring horizontal and vertical orientations, to make

feature groups along with location of the feature group in a local coordinate system attached to the object are used as feature sets. AdaBoost is used to build a strong reliable classier for detection of each part. For a given image, a scale-space pyramid is built and the described features computed at each scale. The different strong classifiers, learned using the AdaBoost framework, of each body part are used to detect the presence of their respective body parts. Given the locations and magnitudes of local maxima provided by individual detectors, a likelihood model is used to combine the detection results. The overall system achieved an 87% detection rate with a 1 false positive per 1.8 images on 400 images taken from the MIT Pedestrian Database 2 and took less than 10 seconds on a 2GHz P4 machine for a 640×480 image.

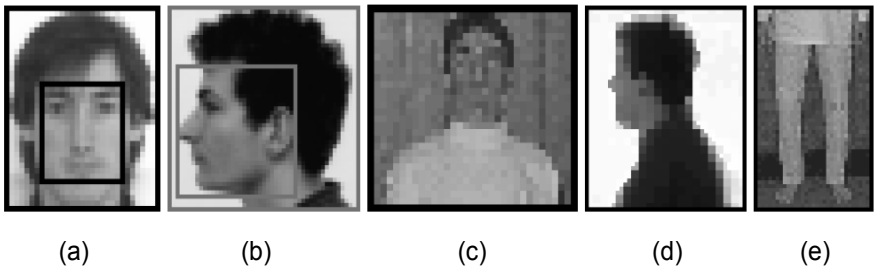


Figure 2.11: [Mikolajczyk 2004] body parts. (a) Frontal head and face (inner frame). (b) Profile head and face (inner frame). (c) Frontal upper body. (d) Profile upper body. (e) Legs

In [Felzenszwalb 2010], a person detection scheme in a general object detection framework with discriminatingly trained parts based models is presented. The person detector is based on mixtures of multi-scale deformable parts models that have the ability to represent a highly variable object class like the one of a person. Used features consists in Histograms of Orientation Gradients (HOGs) with analytically reduced dimension. All model parameter learning was done by constructing a latent SVM problem and training the latent SVM using a coordinated descent approach (see figure 2.12).

In general body parts detection approaches are better suited during occlusions than full body approaches as their detection depends not only on the whole body but on the different parts of the body, head, torso, legs, detected.

In [Yang 2012], a new method for articulated human detection and human pose estimation in static images is proposed. It is based on a new representation of deformable part models. Rather than modelling articulation using a family of warped (rotated and foreshortened) templates, a mixture of small, non-oriented parts are used. The model describes a general, flexible mixture model that jointly captures spatial relations between part locations and co-occurrence relations between part mixtures, augmenting

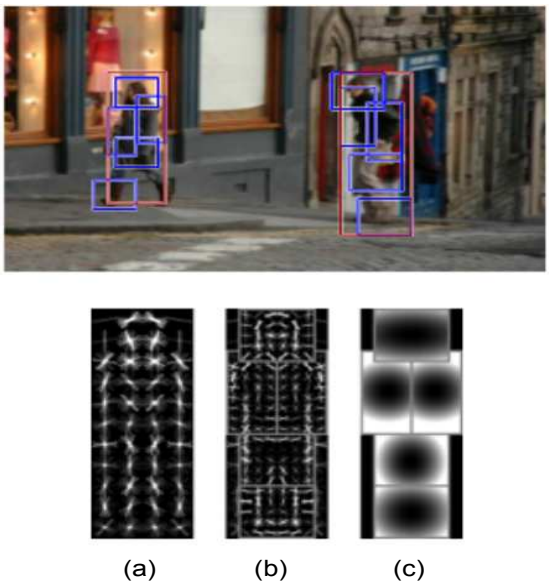


Figure 2.12: Detection obtained with a single component person model ([Felzenszwalb 2010]). The model is defined by: coarse root filter (a), several higher resolution part filters (b) and a spatial model for the location of each part relative to the root (c). The filters specify weights for histogram of oriented gradients features. Their visualisation show the positive weights at different orientations. The visualization of the spatial models reflects the "cost" of placing the center of a part a different locations relative to the root.

standard pictorial structure models ([Felzenszwalb 2000, Ronfard 2002]) that encode just spatial relations (see figure 2.13). All parameters, including local appearances, spatial relations, and co-occurrence relations (which encode local rigidity) are learnt with a structured SVM solver.

2.1.5 Discussion

In this section, we have presented the several existing approaches for people detection in images. The most used scheme fort people detection consists in two main steps: classifiers training, and detection task. Both of these steps require to represent image region information in a significant and useful way. Many features have been used for this purpose. Some of them are specifically designed for object/people detection, like HOGs, Haar-like and Region Covariance Descriptors, and are used for other purpose later, while other features were proposed initially for different tasks, like LBP for texture analysis and Shape Context for shape matching, and have been adapted or added as additional information for people detection task.

The selection of more significant features for the classification is performed using

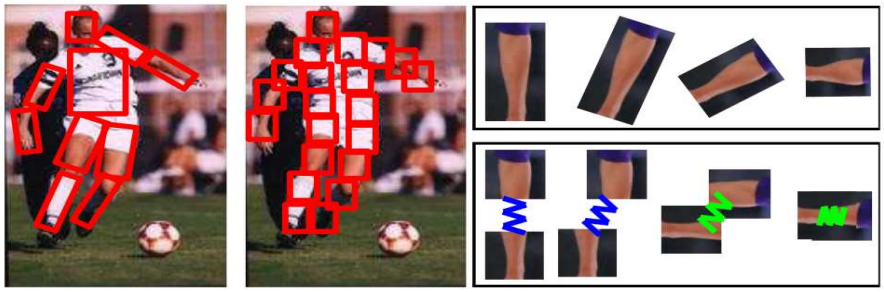


Figure 2.13: [Yang 2012] flexible mixture-of-parts model (middle) differs from classic approaches (left) that model articulation by warping a single template to different orientation and foreshortening states (top right). Small warps are approximated by translating patches connected with a spring (bottom right). For a large warp, a different set of patches and a different spring are used. The proposed model captures the dependence of local part appearance on geometry (i.e. elbows indifferent spatial arrangements look different).

several techniques. The most popular techniques in people detection state of the art relies on machine learning techniques. Two kinds of machine learning stand out from all other by their popularity in people detection state of the art. Support Vector Machines and Adaboost regroup the most proposed training-based approaches.

Final boosting classifiers can be single strong classifiers or cascade of partial strong classifiers. The first type encodes the separation between two classes in a unique structure containing a large set of weak classifiers (the number of required weak classifiers is directly proportional to the separability level of the data), and requires the computing of its whole weak classifiers to provide a classification response for each tested data. It requires constant evaluation time for each candidate data.

On the opposite, the second type encodes the classes separation in successive partial strong classifiers and performs in a rejection scheme, stopping classification evaluation at the first level of cascade which returns a negative response for the tested data. Each cascade level contains significantly less weak classifiers in comparison to a single structure strong classifier for the same data separation. This is due to the role of the considered cascade level which is supposed to separate the class of interest from a subset of the other class only. The candidate data evaluation time is variable and depends on the number of evaluated cascade levels before the rejection (if negative data) or acceptance (if positive data).

Due to the complex separability of human and non-human classes, cascade of classifiers seems to be more adapted and more efficient in term of required processing time as long as their structures are optimal. They have to ensure a high rate of non-human data rejection in the first levels, requiring less weak classifiers evaluation than for a single

strong classifiers. Only small proportion of hard negative data has to require the whole cascade evaluation.

People detection approaches either use a full body or body parts detection approaches.

An important advantage of the part based approach is its better robustness to partial occlusions than the standard approach considering the whole object, due to its object parts segmentation and their independent detection. Nevertheless, part detection methods present two main issues in comparison with full body based approaches, depending on the way they are performed and the resolution of images in which the detection is performed.

First, for the same type of classifiers trained on the same set of features, but one on full body and the other one on body parts, it is clear that body part approaches require more processing time. A full body detector will perform the classifier once on a candidate region while part based detectors require as many classifier applying as considered number of parts. This processing time is more important when no topological constraints are applied to restrict specific body part searching (topological constraints are generally used after detection to infer the existence of a person or not).

Second, according to the defined body part sizes, and the resolution of images, some parts are hard to detect if their corresponding regions in image are too small or have low resolution. We have seen than some proposed approaches ([Yang 2012]) use a large number of very small body parts to ensure a better flexibility of the built model. It may cause miss-detection of some parts and increase processing time. Most of proposed body part based approaches using small part detection, have been tested and evaluated on images in which, people have sufficient image sizes and resolution. This constrains can not be ensured in all situations. The approaches which were evaluated on low resolution images use larger and fewer parts segmentation, loosing some of the effectiveness of using more and smaller parts (for example, [Bak 2010] on iLids dataset, with 5 parts: head, torso, arms and legs).

From these observations, we can deduct that feature selection as well as detection strategy (full body VS. body parts) are dependent on the training data parameters (images resolutions and quality) and final addressed task requirements (real-time constraint for live processing VS. off-line processing).

Due to the addressed problem context, which is video surveillance, and its constraints (low resolution images and small person sizes most of the time), we have decided to focus our work on full body detection approaches. We have used cascade of classifiers structure to reduce decision time (which is a critical point for live surveillance systems) by proposing an efficient approach to optimize the structure of the cascade and to reduce greatly training and detection times.

2.2 Mono-camera Object Tracking

Object tracking is a critical task for scene event understanding. Detecting objects of interest on independent images allows to know that these objects are present in the scene, but does not help to understand what they are doing. In computer vision systems, all kinds of event and behaviour detections require temporal information as long as an event or a behaviour has a starting and ending time (a duration) and consists in successive states. To reason at the level of scene events, it is necessary for a computer vision system to be able to continuously associate the same identifier to the detected object of interest.

The aim of an object tracking algorithm is to ensure this "object of interest"-"unique identifier" association as long as this object is present in the observed scene, and thereby, provide the full trajectory of the object in the scene. This is done by locating its position in each video frame.

Object modelling plays a crucial role in visual tracking because it helps to characterize the object of interest. The modelling is performed in a different sense than for people detection purpose. For people detection the model is a class model. It has to be as generic as possible to cover all possible variations between persons, size and like pose differences, clothing types, etc. while being discriminative against other object classes. In object tracking task, the model is an individual model. It means that an object model has to be full-discriminative against both other type objects as well the ones of the same type. The way in which object modeling is performed can be considered as a first criteria for object tracking approach classification.

Once a model of an object of interest is computed, it is used to track this object by searching it through out the video frames. The model searching/matching over time methods constitute a second criteria for object tracking method classification.

We first present the object modelling methods for tracking. Then, we present object model searching over time methods.

2.2.1 Object Modelling

For object tracking purpose, a wide variety of object modelling approaches have been proposed, depending on the amount and the type of information which are extracted from the object. At the low level, the object can be represented simply by intensity value of its pixels [Pahlavan 1992]. At the middle level it can be represented by some features used independently like, color [Pérez 2002, Comaniciu 2000] or feature points

[Yao 1995, Tissainayagam 2003]. At the highest level it can be represented by a global feature vector which can be boosted from many features to perform a better representation of the object, like [Chau 2011, Zhou 2006] (distance, area, shape ratio and color histogram) or [Serby 2004] (interest points, straight and curved edges, textured regions and homogeneous regions).

In the following paragraphs, some object modelling features are presented.

2.2.1.1 Color Modelling:

Color is one of most fundamental feature to describe an object, due to its strong descriptive power. RGB color space is usually used to represent images; however, the RGB color model is perceptually not a uniform color model. Some other color spaces are more adapted for intra-color distance computation. HSV (Hue, Saturation, Value) and HSL (Hue, Saturation, Lightness) is an approximately uniform color spaces and used intensively in literature.

The color information of object can be modeled in several ways. we can cite:

- **Color histograms**: Color histogram is the most used color encoding method for object modelling [Pérez 2002, Comaniciu 2000, Qian 2007]. It is a representation of the distribution of colors in an image, i.e. it represents the number of pixels that have colors in each of a fixed list of color ranges (bins). It provide an interesting information about the frequency of each color in the object image, but present the inconvenient to do not encode the spacial distribution of the colors. To deal with this issue, some other color encoding methods, which keep color spatial repartition information, have been proposed. We can cite color spatiograms [Birchfield 2005] and Maximally Stable Colour Regions for Recognition and Matching [Forssén 2007].

- **Color spatiogram (for spatial histogram):** Spatiograms capture both occurrences and spatial repartition of colors. In [Birchfield 2005], a second-order spatiogram of an image I is defined as:
  $h_I^{(2)}(b) = \langle n_b, \mu_b, \Sigma_b \rangle \qquad b = 1,\ldots,B$
  where $^{(2)}$ is related to the second order of spatiogram (the proposed spatiogram can be generalized to higher orders), B is the number of used bins, n_b is the number of pixels whose value is the one of the b_{th} bin, and μ_b and Σ_b are the mean vector and covariance matrices, respectively, of the coordinates of those pixels. A comparison between image generation from color histograms and from color spatiograms is provided by authors in figure 2.14.

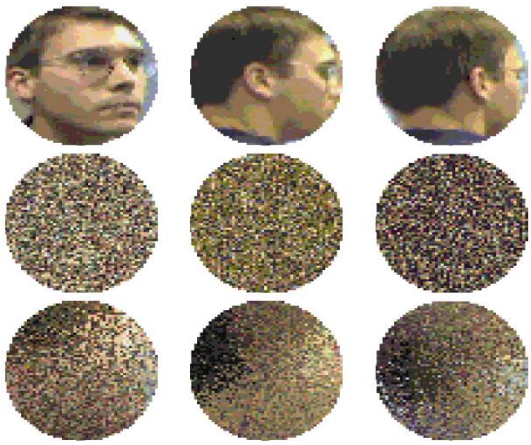


Figure 2.14: Three different poses of a person (top), with images generated from the histogram (middle) and spatiogram (bottom). The spatiogram captures spatial relationships among the colors, whereas the histogram discards all spatial information ([Birchfield 2005]).

- **Maximally Stable Colour Regions (MSCR):** In [Forssén 2007], extends the Maximally Stable Extremal Regions (MSER) [Matas 2002] method to color images. Color regions are iteratively clustered until stable color regions are formed, using predefined thresholds. This method provide a feature vector containing: the color region area, the coordinates of the is color region centroid, the first and the second order moments of the color regions according to the two axis (X and Y). An example MSCR extraction from a color image is provided in figure 2.15.

- **The Dominant Color Descriptor** allows specification of a small number of dominant color values as well as their statistical properties like distribution and variance. Its purpose is to provide an effective, compact and intuitive representation of colors present in a region or image.

- **The Scalable Color Descriptor** is derived from a color histogram defined in the Hue-Saturation-Value (HSV) color space with fixed color space quantization. It uses a Haar transform coefficient encoding, allowing scalable representation of description, as well as complexity scalability of feature extraction and matching procedures.

- **The Color Structure Descriptor**: is also based on color histograms, but aims at identifying localized color distributions using a small structuring window. To guarantee interoperability, the color structure descriptor is bound to the Hue-Min-Max-Difference (HMMD) color space.

- **The Color Layout Descriptor**: captures the spatial layout of the dominant colors on a grid superimposed on a region or image. Representation is based on coefficients of the Discrete Cosine Transform (DCT). This is a very compact descriptor being highly efficient in fast browsing and search applications. It can be applied to still images as well as to video segments.

Figure 2.15: Example of Maximally Stable Color Regions (MSCR) extraction using [Birchfield 2005] method. (a) The original image. (b) An elliptical approximation of the extracted regions

2.2.1.2 Shape Modelling:

The shape features are used as a powerful cue to track object in video frame sequences. Several representation are used (see figure 2.16):

- Point Representation: In visual object tracking, the trivial shape is the point. An object is represented with a pixel location representing either some statistics on the object shape, such as the centroid and , or a particular characteristic of interest like the centred contact point with the floor. Many approaches use single or multiple points to represent tracked objects. For instance, it has been used for in vehicle tracking [Kanhere 2008] or extended objects, such as ships or a convoy of vehicles moving in urban environment tracking [Angelova 2008] and in [Gustafsson 2002] for automotive and airborne tracking applications. In [Chen 2007a], A point representation is used to represent the variance of pixels in the object of interest shape. Points are also used in the calculation of optical flow: due to the large number of vectors to estimate, only the point representation can be afforded [Kragik 2000].

- Primitive geometric shapes: The point representation of an object is a simple model. However, it does not grasp the entire dynamics of the object. To rem-

edy this gap, more advanced parametric shapes are necessary. The most popular parametric shapes are primitive geometric shape (rectangle, square, ellipse and circle). The rectangle representation is frequently used for tracked objects representation like for vehicles tracking [Melo 2006, Chen 2003] or people tracking [Yang 2005]. An adaptive square shape has been used for object representation in [Bradski 1998]. The ellipse offers the advantage of "rounding" the edges compared to the rectangle when the object does not have sharp edges [Chang 2005]. In [Comaniciu 2000, Comaniciu 2003], the author used an elliptical shapes to represent the moving object.

- Articulated shape models: Articulated shapes are employed for tracking if different portions of the object of interest are to be described individually (e.g. legs, arms and head). This kind of representation is much suitable for a human body, which is an articulated object with head, hands, legs etc. These body parts should be linked by a kinematic model. The parts can be represented by any primitive geometric shape such as rectangles, circles and ellipses. In [Ramanan 2003], an articulated shape model, describing the body configuration and disambiguating overlapping tracks, is developed.

- Skeletal models: In this representation a skeleton of object is extracted to model both articulated and rigid objects. A skeleton consists in as a set of articulations within an object that describes the dependencies and defines constraints between the representations of the parts. In [Ali 2001], the skeletal model is used for automatic segmentation and recognition of continuous human activity.

- Object silhouette: The silhouette,also called "Blob", is a dense, non-disjoint, binary mask that represents an object of interest. Blobs are of particular importance for pixel-wise processing. For instance, background subtraction provides blobs identifying the foreground or the moving objects in a scene [Wren 1997], [Stauffer 1999, Elgammal 2002].

- Contour: In this representation the boundary of an object is defined as a contour. A non-rigid object shape can be better represented by these representations [Yilmaz 2004].

2.2.1.3 Texture Modelling:

Texture is another important information for object characterization. It provides a useful information about the spatial arrangement of color or intensities in an image or

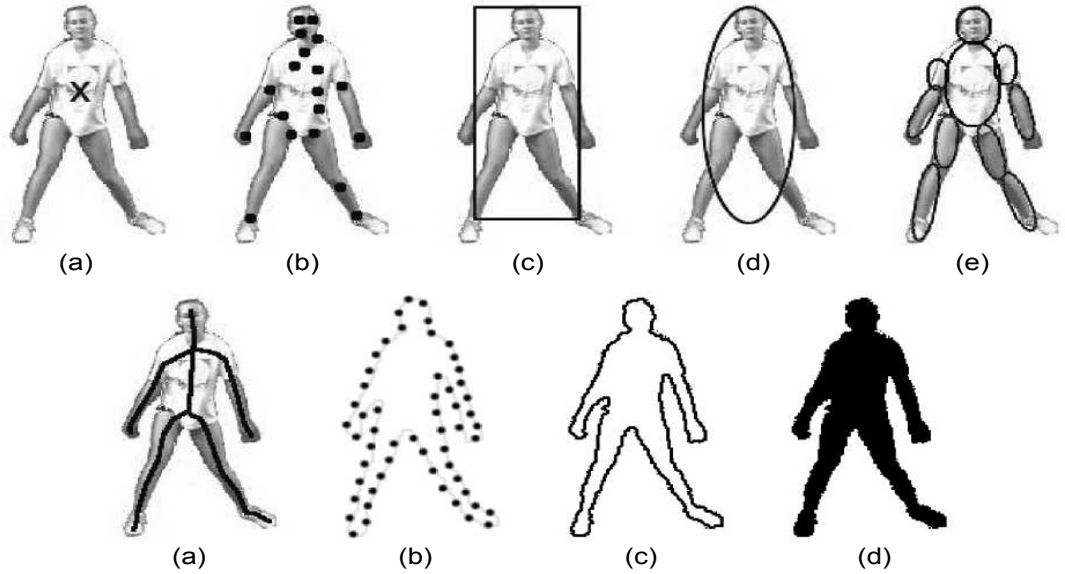


Figure 2.16: Object representations. (a) Centroid, (b) multiple points, (c) rectangular patch, (d) elliptical patch, (e) part-based multiple patches, (f) object skeleton, (g) complete object contour, (h) control points on object contour, (i) object silhouette. (source [Yilmaz 2006])

selected region of an image. Texture is concerned with representing regular patterns in an image [Forsyth 2002]. The texture representations can be performed in several ways, providing different features. These features may differ in the described region sizes, the main extracted information or in the level of robustness and invariance to rotations and illumination changes. Upon the several existing texture features, we can cite Local Binary Pattern (LBP) and Haar-like features (please refer to sections 2.1.1.1 and 2.1.1.4 for more details concerning this two texture features), Co-occurrence Matrices, Edges and Gradient-based local descriptors.

- **Co-occurrence Matrices:** also referred to as GLCM (Gray-Level Co-occurrence Matrices) capture numerical features of a texture using spatial relations of similar gray tones. Numerical features computed from the co-occurrence matrix can be used to represent, compare, and classify textures. The following are a subset of standard features derivable from a normalized co-occurrence matrix:

  Angular 2^{nd} Moment $= \sum_i \sum_j p[i,j]^2$

  Contrast $= \sum_{n=0}^{Ng-1} n^2 \{\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p[i,j]\}, \qquad$ where $|i-j| = n$

  Correlation $= \frac{\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} (ij)p[i,j] - \mu_x \mu_y}{\sigma_x \sigma_y}$

$$\text{Entropy} = -\sum_i \sum_j p[i,j] \log(p[i,j])$$

where $p[i,j]$ is the $[i,j]^{th}$ entry in a gray-tone spatial dependence matrix, and Ng is the number of distinct gray-levels in the quantized image.

For object tracking purpose, Chen et al. [Chen 2009] construct Kernel Co-occurrence Matrices (KCMs) to represent the target model and the target candidates. Those matrices are employed as the tracking cues in mean shift framework. The angle relation between pixel-pairs is redefined to depict the asymmetric characteristic of the objects. The KCMs of the target model and the candidates are normalized to a same integer to increase calculation accuracy. The computation of each pixel weight is modified to improve operation speed. The tracking results of several real world sequences with dark illumination or lighting variance show that the proposed algorithm can track the target effectively.

- **Edges:** Edge detection aims to identify image pixels with brightness discontinuities, i.e. pixels at which the image brightness changes sharply. The points at which image brightness changes sharply are typically organized into a set of curved line segments named edges.

  Edge detection is performed by applying specific operators on image. Many edge detection operators have been proposed. We can cite Sobel Operator, Robertś cross operator, Laplacian of Gaussian, and Canny edge detection algorithm which is known to many as the optimal edge detector (see figure 2.17). The main difference between the operators resides is they handle different edge orientations.

  Edges have been used for people tracking in some approaches. In [Murshed 2011], Canny edge map is used to characterize moving object region. Curvature-based features are used for moving edge registration due to its transformation invariance nature. Each individual edge segment is tracked using a Kalman filter. Edge segments are clustered by using a weighted mean shift algorithm. The final moving object tracking is performed using a group motion tracker, applied on each cluster. Due to the robustness of edges against illumination changes and partial occlusions, the proposed tracked performs efficiently.

  In [Zhu 2006a], an edge-based tracking algorithm is proposed. The feature points are extracted by efficiently utilizing the image edges in the object region. Then the parameter vector of the object's motion model is estimated based on minimizing the sum-of-squared differences between the reference feature points in the reference frame and the observed feature points in the tracking sequence frame. The

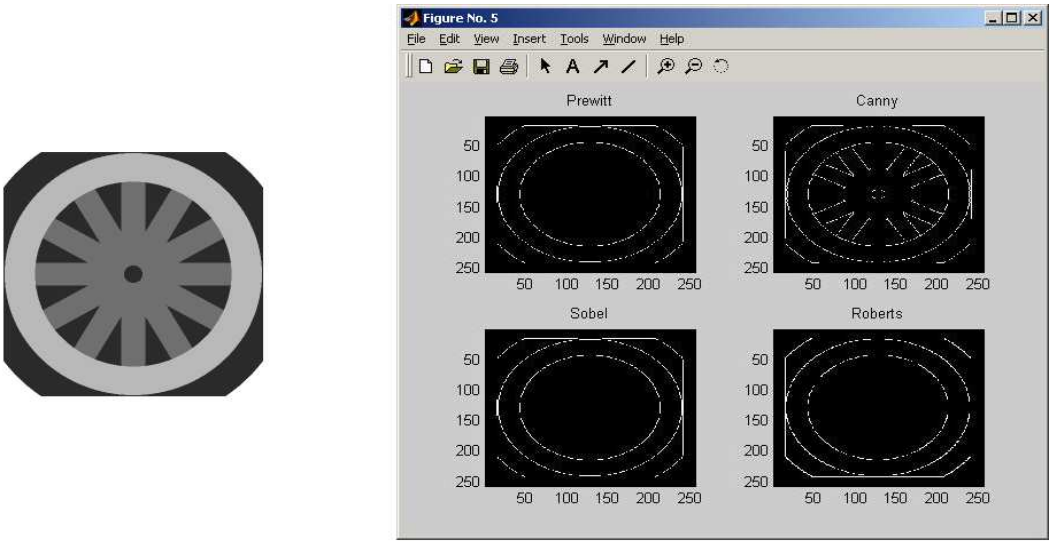


Figure 2.17: Comparison between edge detectors. (a) A test image. (b) Results of edge detection on Figure 7. Canny had the best results (source [Maini 2009]).

experiments show that the edge-based tracking algorithm proposed by us can track object efficiently under uniform and varying illumination conditions.

- **Gradient-based local descriptors**: Many gradient-based local descriptors have been proposed to characterize the textures of an image region. SIFT (Scale Invariant Feature Transform) [Lowe 2004] (see 5.2.1 section for more details) and its derivative (PCA-SIFT, GLOH, DAISY, etc.) and HOG (Histograms of Oriented Gradients) [Dalal 2005] (see section 2.1.1.3) are the most known and the most used for tracking.

  SIFT and HOG descriptors perform with similar principle in the sense that both take the raw information from the magnitudes and orientations of pixel gradients in the described region of image to build histograms. The main difference consists in:

  - The way the raw information (gradient magnitudes and orientations) is processed: different weights for pixel contributions, according to their position with respect to the centre of considerer region, are used for SIFT descriptor computing while similar weight for all pixel contributions are used in a HOG cell.
  - The way the information is encoded: gradient histograms are extracted from

independent pixel regions for SIFT, requiring and independent normalization to ensure robustness against illumination changes, while HOG use overlapped block of pixels for computation, which implicitly ensure the local invariance to illumination changes

– The robustness to 2D rotations: SIFT descriptor is invariant to 2D rotations because it is computed using relative gradient orientations with respect to the main gradients orientation in the considered region. HOG descriptor does not integrate a 2D rotation invariance and is computed using absolute gradients orientations in the image region.

Note that these two descriptors were initially proposed to perform using gray scale images. Some variants SIFT descriptor integrating color information have been proposed, like in [Verma 2011] and [Abdel-Hakim 2006] using RGB color space, and [Bosch 2006] where HSV color space has been used.

These two descriptors were used for object tracking in a large amount of approaches.

In [Zhou 009], a SIFT based mean shift algorithm is presented for object tracking. SIFT features are used to correspond the region of interests across frames. Meanwhile, mean shift is applied to conduct similarity search via color histograms. The probability distributions from these two measurements are evaluated in an expectation-maximization scheme so as to achieve maximum likelihood estimation of similar regions. This mutual support mechanism provides better tracking performance if one of the two measurements becomes unstable.

In [Fazli 2009], A color SIFT based particle filter algorithm is proposed for object tracking. It provides interesting tracking performances in partial occlusions, rotation and scale variations conditions.

2.2.1.4 Motion Modelling:

Motion detection is a critical part of the human vision system [Sonka 2007]. Optical flow is the most widespread depiction of motion. Optical flow represents motion as a displacement vectors which defines the movement of each pixel in a region between subsequent frames [Horn 1981]. Horn and Schunk [Horn 1981] computed displacement vectors using brightness constraint, which assumes brightness constancy of corresponding pixels in consecutive frames. Lucas and Kanade [Lucas 1981] proposed a method that computes optical flow more robustly over multiple scales using a pyramid scheme.

Optical flow is used for object tracking in some approaches. In [Shin 2005], the proposed tracking algorithm performs in three steps: First, the object of interest is localized using optical flow detection. Then prediction and correction of the object's position is performed using spatio-temporal information of the optical flow computed on beforehand detected feature points. Finally restoration of occlusion using an introduced non-prior training (NPT) active feature model (AFM) framework. The proposed algorithm can track both rigid and deformable objects, and is robust against the object's sudden motion because both a feature point and the corresponding motion direction are tracked at the same time. The authors claim that the proposed AFM enables stable tracking of occluded objects with maximum 60% occlusion.

In [Chen 2011], object contours are tracked with optical flow. This algorithm achieves accurate, rapid and stable object tracking on the evaluated situation.

Optical flow is an interesting way to model moving object as long as the addressed situations are not complex, like the conditions in which [Shin 2005] and [Chen 2011] were evaluated. Single moving object in the scene or collusions resulting from uniform object movements can be well handled. In more complex scenes with multiple moving objects with non-uniform movements, this technique may fail.

2.2.2 Object Tracking Techniques

Once an object of interest is modelled, the next step consists in its localization in each frame of the video sequence. Tracking methods can be divided into two general types: deterministic and probabilistic methods. Deterministic methods look for the local maxima of a similarity measure between the object model and the considered candidate regions of current image while probabilistic methods aim at modelling object movement and to perform the object localisation through successive state prediction/update steps.

2.2.2.1 Deterministic Methods

A deterministic systems is a system in which no randomness is involved in the development of the future states. Deterministic tracking methods always provide the same tracking results on the same input data (video sequences and parameters). In this kind of approaches, the object localization may be performed either by an exhaustive search of the object model on the whole image (least-square tracking), by an iterative neighbour exploration (Mean-shift tracking), by point matching, or by many other methods.

- **Least-square tracking:** Least-square tracking consists in an exhaustive search of the tracked object model in the whole image. A Similarity measure between the object model and candidate models extracted from all image positions is computed.

The position of the candidate model which provides the best similarity measure is considered as the new localisation of the tracked object.

Least-square tracking methods, applied on the whole image, are the most effective tracking methods in terms of correct localisation of tracked objects on video sequences since they explore all the possibilities space. However, these methods are also the most time consuming among all tracking methods. The required processing time is unacceptable for most of applications. The evaluation of the model matching in all possible positions requires many computations, especially if the model building and the similarity measure computing are complexes to perform (multi-features model for example). To avoid this critical issue, two main heuristic solutions can be used.

The first one consists in the reduction of the possibility space by performing the exhaustive search in a "local" region instead of the whole image. This really decreases the computation time in relative proportion to the region reduction, leading to another difficulty: "local" is an ambiguous term. It is hard to define an optimal local region for exhaustive search without any a priori knowledge on the object motion. Bad local region localization automatically induces a tracking failure if the correct matching region is not included.

The second heuristic solution consists in a progressive search. Instead of testing the whole possibility space, a uniform browsing step is used to evaluate a subset of positions. For example, a searching process with 2 pixels step in both directions divide the number of the tested possibilities by 4, which is not a negligible factor. The more important the step is, the lesser the number of matching tests are. However, this solution, as the previous one, presents an important issue: how to choose the optimal browsing step? Object feature sensitivity to shifting may greatly vary. Some features like color histograms are less sensitive to small shifting (due to the contained information nature, i.e. color frequency and non-spatial information) than other features like Region Covariance Descriptors built with spatial information. For more sensitive feature-based object models, the maximum similarity measure may not correspond to the real object but to another image region, if the browsing process does not test the real object region.

- **Mean-shift tracking:** Mean-shift tracking assumes that the object locations in successive images are close to each other. Thus, the searching process in the current image starts from the same location as the object location in the previous image and is performed iteratively. At each iteration, the direct neighbourhood of the current mean-shift window location is tested. A similarity measure is performed

between the tracked object model, and those of candidate region located in neighbouring positions. The location with the best similarity value is taken as the new position of the mean-shift window for next iteration. The process is repeated until it stabilizes, i.e. all the similarity values of all neighbouring positions are less than the one for the current mean-shift window position. Mean shift-tracking seems to be an efficient approach, but it presents two main issues.

First, the assumption of small object displacements in successive frames can not be ensured. If the mean shift window is not correctly initialized (smaller than the object displacement), the tracking will probably fail.

Second, the convergence time is very variable and depends mainly on the object displacement magnitude and on the type and distribution of the features on which the object model has been built.

Many deterministic approaches have been proposed in the state of the art.

In [Chen 2009], moving object models are constructed using Kernel Co-occurrence Matrices (KCMs). Then those matrices are employed as the tracking cues in a mean shift framework. In [Birchfield 2005], proposed spatiograms are used as object model for tracking. They are tracked by a proposed kernel-based tracker deriving from mean shift method. Another mean-shift based approach is proposed in [Comaniciu 2000] to track non-rigid objects. The dissimilarity between the target model (its color distribution) and the target candidates is expressed by a metric derived from the Bhattacharyya coefficient. In [Zhou 2006], a Mean-shift method is used to track SIFT [Lowe 2004] features and color histograms, which are to model the tracked objects. Finally, in [Thayananthan 2003], a derived mean-shift method is applied to track human hand using its shape. The hand is modelled with Shape Context feature [Belongie 2002] (see section 2.1.1.5) and the used similarity measure is the Chamfer Matching method [Barrow 1977] (see section 2.1.3.1)

Previously mentioned issues for mean-shift tracking methods can affect the tracking performance since they were not handled in this proposed approach.

In [Zhu 2006a], a multi-feature points correspondences approaches is proposed. Feature points are extracted on the image edges in the object region. Then the parameter vector of the object's motion model is estimated based on minimizing the sum-of-squared differences between the reference feature points in the reference frame and the observed feature points in the tracking sequence frame. The experiments show that the edge-based tracking algorithm proposed by us can track object efficiently under uniform and varying illumination conditions.

In [Bilinski 2009], another point matching based method is proposed. FAST points detector [Rosten 2006] is applied to detect points of interest on moving objects, and this on each frame. Each detected point is associated to a HOG descriptor [Dalal 2005] computed around it. The tracking is then performed by a direct point matching between object of previous frame with those of the current frame, by computing their HOG descriptors similarity. The final object movement is determined using the trajectories of all matched points of interest.

The point matching based methods provide good tracking performances, but present the main issue of high computational time consuming. These approaches requites to repeat two non-basic tasks which are interest point detection, and descriptors computation. In addition to that, the matching process consist in a kind of Cartesian product matching, i.e. each interest point from the object to track is compared to all interest points of the candidate object by their associated descriptor similarities. It is a slow task. This last issue can be partially handled by using some localisation constrains. For example, points belonging to the upper part of the object to track will be compared with those of the upper part of the candidate object only, and not with all the points. This solution decrease matching time but can cause matching fail in case of deformable or partially occluded objects.

2.2.2.2 Probabilistic Methods

They consist in a recursive estimation of a hidden state of a moving object using noisy observations. Considering that the hidden state evolves over the time, it is necessary to introduce an *a priori* model of displacement for the mobile, and to consider the estimation problem in a Bayesian framework.

According to the types of motion and noise, different kinds of Bayesian filters are defined. Two main Bayesian filters are used for object tracking. In the particular case of Gaussian linear systems, the filtering problem has an explicit solution, called **Kalman Filter**. In the case of non-linear systems with noise which is not necessary a Gaussian one, or in the general case of hidden Markonv models, some very effective Monte Carlo methods have appeared under the name of **Particle Filters**. Intuitively, each particle represents a possible state, explores the state space following the *a priori* motion model and is duplicated or eliminated at the next generation depending on its coherency with the current observations, quantified by likelihood function. This mutation/selection mechanism has the effect of automatically concentrating the particles in the regions of interest inside the state space.

■ **The Kalman Filter**

As it was indicated before, Kalman filter allows to estimate the parameters of a system which evolves over the time, using some noisy measurements.

The Kalman filter is performed in two successive steps:

– Prediction: The first step is the prediction of the estimation according to the system model. To perform it, the Kalman filter takes the previous estimation of the parameters and the error, and uses them to predict the new parameters and the error depending on the system model.

$$X_t^- = DX_{t-1}^+ + W \tag{2.2}$$

$$P_t^- = DP_{t-1}^+ D^T + Q \tag{2.3}$$

where:
X_t^- and X_{t-1}^+ are respectively the predicted and corrected states at time t and t-1; P_t^- and P_{t-1}^+ are respectively the predicted and corrected covariances at time t and t-1; D is the state transition matrix which defines the relation between the state variables at time t and t-1; W is a noise matrix and Q is its covariance.

– Correction: The second step updates this prediction thanks to the new measurements Z_t. These measure (which are noisy) allow to obtain an estimation of the parameters and the error from the performed prediction. If the model contains errors, this update step allows to correct them.

$$K_t = P_t^- M^T [MP_t^- M^T + R_t]^{-1} \tag{2.4}$$

$$X_t^+ = X_t^- + K_t[Zt - MX_t^-] \tag{2.5}$$

$$P_t^+ = P_t^- - K_t M P_t^- \tag{2.6}$$

where M is the measurement prediction matrix, K is the Kalman gain and R is the covariance matrix of measurements noise.

Kalman filter has several interesting aspects. The power of this filter lies in its ability to predict the parameters and to correct errors, not only the sensor ones but also those of the model itself.

In fact, to apply a Kalman filter to estimate the parameters of a given system, it is necessary to provide a linear model before. Some variants of the Kalman filter, to deal with non-linear models, have been proposed, for example the Extended Kalman Filter which is discussed below.

In a classical estimation method, like least squares method, a simple error in the system model inevitably leads to an error in the estimation. The advantage of Kalman filter is to integrate a term of imprecision on the model itself, allowing it to correct the estimation instead of model errors (Of course, the model error has to be reasonable).

Another advantage of the Kalman filter is its ability to determine the mean error of its estimation. In fact, Kalman filter provides a vector of estimated parameters, but also the error covariance matrix. This matrix informs us about the precision of the estimation. Another interesting information is that the convergence of this error is guaranteed (in case of linear dynamics).

However, this filter is not necessarily the tool to apply in all cases. In fact, as we have seen, we need to model the system precisely to design an efficient filter. The problem is that some systems are hard to model and are not linear.

In the case where the model is a rough approximation, the filter will not be efficient enough to correct the error, which will converge quickly.

To avoid this problem of linear model of the system, the Extended Kalman filter was developed. It allows to deal with non-linear models. However, this method has some defects. First, the error covariance will not necessary converge (unlike the standard Kalman filter for linear models). The second defect is the high computing cost. In fact, some costly new matrices appear in the filtering computation (matrices of partial derivatives of the state equations and measurement of the system model).

Another important limitation of this method is that the Kalman filter allows us to consider only a Gaussian noise model. Noise can generally be modeled as Gaussian, but in some cases, other types of noise can occur. This restriction limits the use of the Kalman filter.

The Kalman filter is an interesting method of estimation, but can only be used when it is possible to accurately model the system and when the noise is a Gaussian one. When it is impossible to find a correct model for the system, it is better to refer to other methods such as Monte Carlo method, called Particle Filter, which is a statistical method, but it requires significant computing power.

■ **The Particle Filter**

Like Kalman filter, Particle filter allows to estimate the parameters of systems which evolve linearly over the time, but also non-linearly, using some noisy measures, even if this noise in not a Gaussian one. It is performed in prediction and correction steps too.

Particle filtering is a global method based on the exploration of the state space of the problem using a set of **particles**. These particles are distributed according to the conditional probability of the process to be estimated constrained by the observations provided by the sensors.

This method does not require the explicit resolution of the equations of the problem, so it is applicable regardless of the complexity of these equations, especially in terms of non-linear and non-Gaussian nature.

In its basic version, a particle filter consists of N particles which evolve in parallel. Each particle evolves according to the measurements provided by the sensors at the sampling time "t" and simulates a possible trajectory, ie. the evolution of a process respecting the same equations than the process to estimate. Each particle provides two information as output:

- A state vector with the same structure than the state vector of the process to estimate.
- A scalar value called *weight*, representative of the probability that this vector is the one of the process to be estimated.

Sequential importance sampling (SIS) is the most basic method used for particle weighting in Particle filters. The weight of each particle is continuously updated over the time. For a given particle, its weight at current time depends on its previous weight, its previous and current state estimations, and the current observation measurements:

$$w_t^i \sim f(w_{t-1}^i, x_{t-1}^i, x_t^i, z_t) \tag{2.7}$$

where w_t^i and w_{t-1}^i are respectively the weights of a given particle P_i at time t and t-1; x_{t-1}^i and x_t^i the estimated state of this particle at time t and t-1 respectively, and z_t is the observation measurements at time t.

This method present the important issue of information degeneration. In fact, only few particles may have a significant weight at each iteration. The particles

corresponding to unlikely hypothesis may continuously degenerate, causing possible divergence of the particle filter, or useless processing of very low weighted particles.

One common way to deal with this degeneracy is resampling. Sampling Importance Resampling (SIR) method consists in removing less significant particles (with the lowest weights) while creating new particles at the same position than the most significant particles (with the highest weights). The number of created particles at the same position of an important particle is proportional to the weight of this important particle. The last step consist in affecting the same normalized weight 1/N to all the particles (N is the number of considered particles) (see figure 2.18).

In this method, the importance of a given hypothesis at the end of an iteration is not represented by its associated particle's weight, but by the number of particles which are related to it. The most important hypothesis at a given iteration will be the basis of more new hypothesis in the next iteration, and information degeneracy is avoided.

For a high enough number of particles, it is possible to demonstrate that the set of all the states of the weighted particles is representative of the conditional probability law of the state vector of the process.

This procedure allows to concentrate the exploratory ability of the network of particles in areas where probability is maximum, thus increasing the precision of the estimate.

The most important limitation of Particle filtering method is its high computation time in comparison with other probabilistic methods (Kalman filter). It is directly proportional to the number of hypothesis to maintain and to process in parallel and to the hypothesis verification (particle weighting) complexity (complex features comparison for example).

For object tracking in video sequences, probabilistic tracking methods have been widely used. In [Elgammal 2002, Comaniciu 2003, Melo 2006, Murshed 2011], a Kalman filter is used to track the several proposed models of object of interest. In [Pérez 2002, Serby 2004, Yang 2005, Fazli 2009], the authors used particles filters for their object tracking approaches.

A summary of all cited state of the art approaches for object tracking, according to their proposed models and their tracking techniques is displayed in Table 2.1;

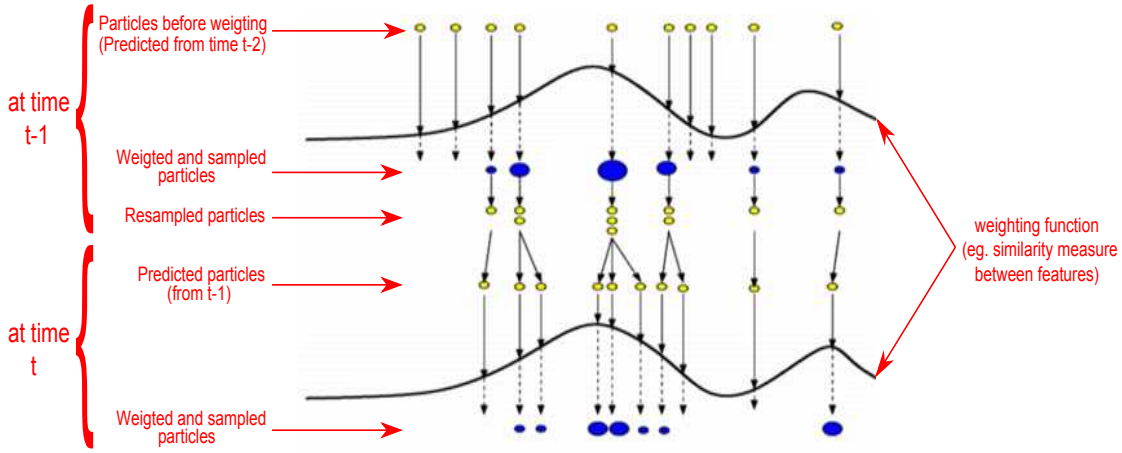


Figure 2.18: Sampling Importance Resampling (SIR) process

2.2.3 Discussion

Due to its importance for scene understanding and useful event information extraction, object tracking has always highly attracted attention of researchers. Many works have been done to deal with the most important challenges (object occlusions, crowded scenes, etc.).

The object modelling is the first important task to be performed. The models have to be highly discriminative to allow a more reliable matching between the same object images over time while avoiding matching errors with a third object. It has to be robust to partial occlusions, illumination changes, and rotations.

Many feature extraction techniques have been proposed in the state of the art. All these features are extracted from one of the main image information families which are texture, color, shape and motion. Several encoding ways have been proposed for each information type, focusing more on some aspect rather than others. For example, color histograms encode the occurrence frequency of each color value without considering the spacial information of the color distribution while The Color Layout Descriptor captures the spatial layout of dominant colors only.

Some state of the art object models are built using a unique feature type while others are built using combination of different features (see Table 2.1). These second kinds of models generally use a concatenation of feature descriptors to provide the final model feature.

Using several concatenated features to build the final object model is an efficient way to improve the discriminative power and the robustness of the object of interest, but it has a cost in term of processing time. Features extraction as well as similarity measure

<table>
<tr><th></th><th colspan="4">Object Modelling</th></tr>
<tr><th></th><th>Texture</th><th>Color</th><th>Shape</th><th>Motion</th></tr>
<tr><td rowspan="4">Probabilistic Tracking</td><td rowspan="3">[Murshed 2011]
[Mörwald 2009]</td><td rowspan="3">[Pérez 2002]
[Qian 2007]</td><td>[Angelova 2008]
[Gustafsson 2002]
[Melo 2006]
[Comaniciu 2003]
[Stauffer 1999]
[Wren 1997]
[Ramanan 2003]
[Chang 2005]
[Yang 2005]</td><td rowspan="4"></td></tr>
<tr><td colspan="2">[Chau 2011]
[Elgammal 2002]</td></tr>
<tr><td colspan="2">[Fazli 2009]</td><td></td></tr>
<tr><td colspan="3">[Serby 2004]</td></tr>
<tr><td rowspan="3">Deterministic Tracking</td><td>[Zhu 2006a]
[Chen 2009]
[Bilinski 2009]</td><td>[Birchfield 2005]
[Comaniciu 2000]</td><td rowspan="3">[Thayananthan 2003]</td><td rowspan="3">[Shin 2005]</td></tr>
<tr><td colspan="2">[Zhou 009]</td></tr>
</table>

Table 2.1: Summary of object tracking approaches, organized by the model information and the tracking technique.

take more time than the computation of a single feature model. In addition, the simple concatenation of feature descriptors provide a decorrelated inter-feature information, which is not efficient since the information is extracted anyway.

This last defect can be handled by using other representation that a simple concatenation. Region Covariance descriptors are a good solution since many image features can be encoded in a single structure which captures not only each feature variation but also correlation between all features. The main disadvantage of this descriptor is its high computation and comparing processing time. A detailed presentation of this descriptor is provided in the chapter 4.

Object modelling is also very dependent on the video properties (color, resolution, compression (noise)) and context (occlusions, illuminations changes, crowded scenes, restricted trajectories, et.). State of the art proposed methods perform well in some conditions while they fail in others.

For all these reasons, and due to the addressed context, which is video surveillance, and its constraints (real-time processing, small objects in large view images, both rigid and non-rigid object tracking, compressed and medium resolution images), we have decided to model our tracked objects using a single feature type which is SIFT feature. This choice is justified by many reasons. First, the real-time processing requirement is better ensured by single feature model. Second, the necessity to have a non-rigid model and to handle partial occlusions are ensured by SIFT point representation. Finally, the robustness of the model to illumination changes, provided by the illumination invariance of SIFT features.

The weakness of single feature models are compensated by our proposed SIFT feature detection, selection, and tracking methods. The other video properties which usually affect tracking performances, like noise (due to image compression) and small object size in images (due to large view of video surveillance camera) are also handled in our proposed method which is detailed in chapter 5.

From tracking technique point of view, the state of the art approaches are divided into deterministic and probabilistic methods. Exhaustive research tracking method, which belongs to deterministic techniques, is the most effective method due to the fact that all the possibilities are evaluated. However, it is not applicable for many applications, especially video surveillance, due to the high processing time it requires. Other deterministic methods such as Mean-shift or direct point matching are interesting methods and provide good tracking results as long as the main conditions are reached (initialization of mean-shift, reduction of point matching possibilities), otherwise, tracking can fail or become very slow.

Probabilistic tracking methods are widely used in object tracking approaches in the state of the art. In opposite to deterministic approaches, probabilistic approaches integrate some randomness in their processing, generally assigned to measurement uncertainty or noise modelling. The same input data may provide different results from a processing to another one.

The main common principle of these methods is to model object motion and to use this model to localize tracked object in two successive steps which are the prediction and the correction (update) one.

The two main used probabilistic tracking methods are Kalman filters and Particle filters. While the Kalman filter tracks an object of interest in a single hypothesis mode, Particle filters use a set of particles to perform tracking. Each particle corresponds to a hypothesis of the system state (mainly tracked object localisation). Particle filters are more interesting since they allow to explore a large space of hypotheses in a parallel way. This advantage constitutes at the same time the most important inconvenient of Particle

filters. In fact, processing a large set of particles, especially during the weighting step which generally requires feature similarity measure computation, can be unmanageable for real-time tracking. However, by using a small but sufficient set of particle, the real-time processing becomes possible.

Due to its previously cited advantages, we have decided to use a particle filtering method for our object tracking. The selected parameters allow our tracking to be performed in real-time.

2.3 People Re-identification

People re-identification in camera network has become a critical task in these last years. With the increase in the number of deployed cameras in restricted and large areas, it has become important to have a global understanding of what is happening in a given location covered by many cameras. Data coming from several cameras should no longer be treated as independent information but rather as global information. Some behaviours or events of interest can only be inferred from a long term tracking of the person of the interest across the camera network (for instance, a person who leaves his/her luggage in an airport and stays more or less near it, is not a suspicious person. On the other hand, if this person goes away from his/her luggage and leaves the airport observed by several cameras, this behaviour should attract attention). Being able to track or find a given person in a camera network on live stream or to localize him/her a posteriori on recorded videos become very important in many applications, especially for security, but not only. Many other applications like marketing for shopping mall and statistics in some sports require to track people on multiple cameras to infer the most frequent shopping paths and thereby to reorganize shops, or to calculate the travelled distance by a football player and the number of passes and shots he has made.

This global reasoning on large areas, under camera networks cannot be automatically performed without robust techniques to maintain the same identity of a given tracked person, regardless of where he/she is located and which camera is observing him/her. This identity maintaining for a given person from a camera to another one is called “re-identification”. Recently, a growing number of studies have been done and this problematic is still attracting more interest from researchers. In the following sections, a literature review of different people re-identification approaches are presented.

According to the kind of used information for re-identification, we can distinguish two main families of approaches: biometric approaches and appearance-based approaches.

2.3.1 Biometric Approaches

Biometric approaches seem to be the most efficient techniques for people re-identification as long as they use biological characteristics of people to identify them. For instance, Pr. John DAUGMAN (Computer vision laboratory of University of Cambridge) estimates that the probability to find two identical irises is approximately $1/10^{72}$ (even for identical twins). The same observation can be done for fingerprint comparison. Sir Francis Galton [Galton 1892] published a detailed statistical model of fingerprint analysis a in his book "Finger Prints". He had calculated that the chance to have two different individuals with the same fingerprints was about 1 in 64 billion.

Unfortunately and as it was mentioned in the first chapter of this thesis, this kind of approach cannot be used in large wide video surveillance systems due to the technical and practical constraints. In fact, this kind of approaches requires specific sensors (for iris and fingerprint recognition), or a high resolution images and sufficient framerate (for faces and gait recognition). In addition to this technical issues, this kind of approaches requires entire collaboration and voluntary actions from people for iris and fingerprint analysis, and depend highly on the orientation of the observed people for face and gait recognition.

In the following sections, we present the state of the art of the most important biometric approaches which are iris, fingerprint, face and gait recognition.

2.3.1.1 Iris Recognition

The concept of iris recognition has been proposed initially in 1936 by the ophthalmologist Frank Burch as a way for people identification. In 1987, Dr. Aran Safir and Dr. Léonard Flom, two ophthalmologists, have patented this idea and in 1989, they asked Pr. John Daugman (Teacher at Harvard University at this time) to develop some algorithms for iris recognition. These algorithms (based on Gabor wavelets) which Daugman have patented in 1994 became the basis for all iris recognition systems. This algorithms are implemented in a system called IrisCode®

In [Daugman 2002], the author explain his algorithm for iris recognition, which is performed in four successive steps. The approach starts by localizing, segmenting and normalizing the iris on image. To capture the rich details of iris patterns, it is necessary to ensure a minimum of 70 pixels in iris radius. Image captures are performed with sensors using NIR (near infra-red) illumination in the 700-900 nm band to be unintrusive to humans. Once iris localized and normalized, its features are extracted and encoded using 2D Gabor filters providing the iris "signature" (See figure 2.19). The dissimilarity measure between two irises is calculates using a fractional Hamming Distance. The last

step consists in the final decision when two irises are compared and their dissimilarity measure is calculated. Generally, in this kind of tasks (check if two data matches), the usual way to take the decision is to define a threshold under which, a dissimilarity measure means that the two data effectively matches. This threshold is generally obtained by a training on a large database of labelled data. Daugman propose another and interesting way to take the final decision for iris recognition. Using the Bernoulli distribution, he predict the distribution between inter-class distances and thereby he fixed optimal thresholds that he generalized for larger datasets without any training dataset.

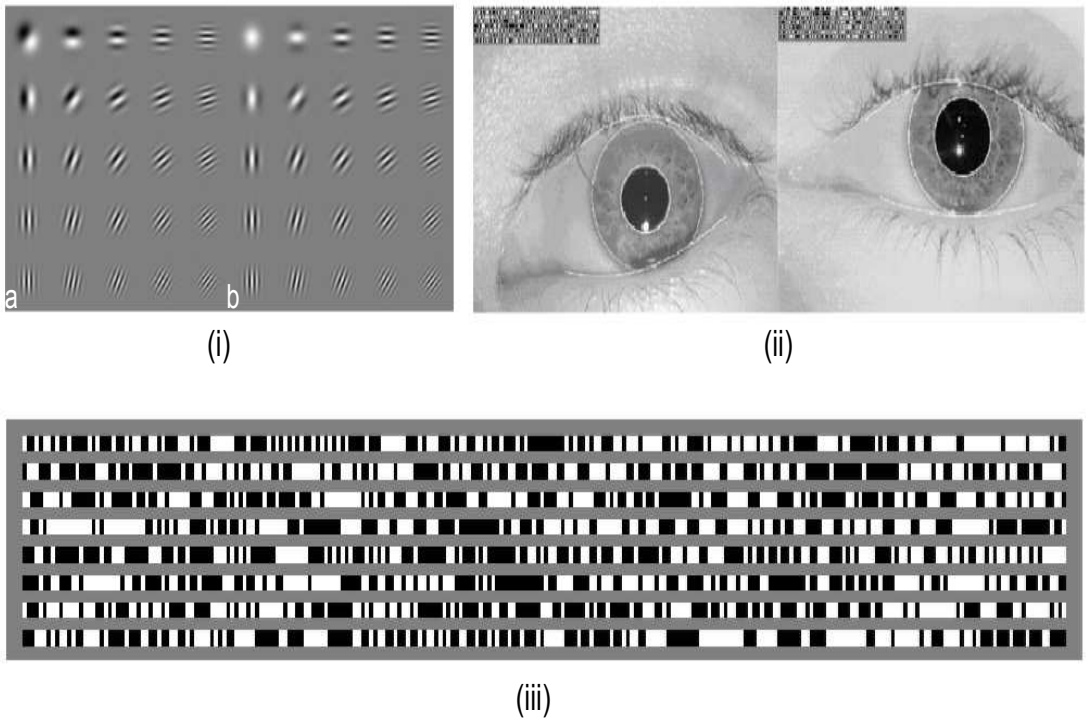


Figure 2.19: Daugman's IrisCode: (i) The set of used Gabor filters with different orientations and different resolution ((a) real parts / (b) imaginary parts). (ii) Localized irises with IrisCode®. (iii) Pictorial representation of IrisCode®.

Wildes [Wildes 1997] proposed an alternative method for iris recognition, keeping the same steps than Daugman, but with different methods for each step. The localization and segmentation of the iris is performed by a circular and elliptic Hough transform, and the filtering is performed by Laplacian of Gaussian filters on four resolutions.

Miyazawa et al. [Miyazawa 2005] introduce the concept of phase correlation for iris recognition. They use a phase correlation based on a band-limited Fourier transform to avoid low quality iris images. They proposed a method to normalize correlation scores

according to the used image size after noise detection.

The IriTech company proposed another patented alternative system to IrisCode® of Daugman. In [Kim 2001], the proposed approach uses Haar wavelets for multi-resolution analysis. 1024 Haar coefficient are computed on different iris zones (See figure 2.20). These coefficients are compared with each other by calculating the difference between coefficients of high frequences and those of low frequences which are not generated by eyelids and eyelashes.

Masek [Masek 2003] and OSIRIS (Open Source for Iris) are two open source reference systems for iris recognition benchmarking. They are built on Daugman and Wildes approaches.

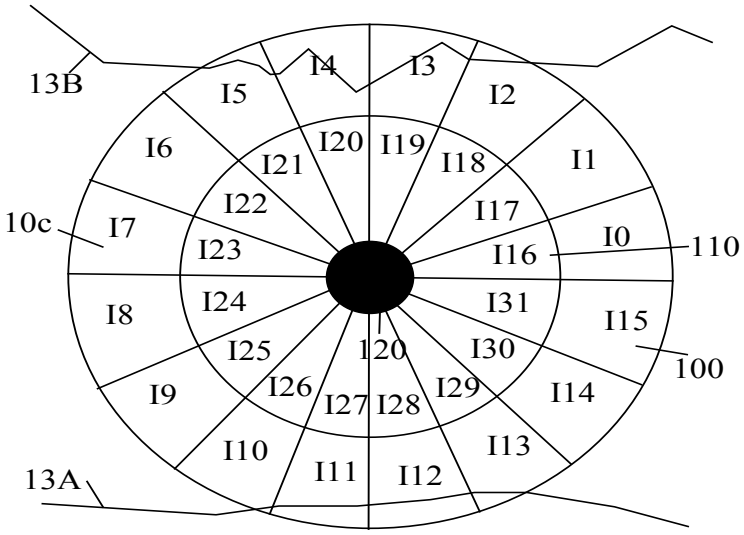


Figure 2.20: IriTech iris subdivision in zones

2.3.1.2 Finger Print Analysis

The Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) defines the fingerprint as an impression of the friction ridges of all or any part of the finger (See figure 2.21(a)). Fingerprint identification, also known as "dactyloscopy", consists in comparing the major features of two fingerprints, called "Minutiae" (See figure 2.21(b)), to determine whether these fingerprints could have come from the same individual. These major features are:

- **Crossover or bridge**: a short ridge that runs between two parallel ridges
- **Core**: a U-turn in the ridge pattern
- **Bifurcation**: a single ridge that divides into two ridges
- **Ridge ending**: the abrupt end of a ridge

- **Island**: a single small ridge inside a short ridge or ridge ending that is not connected to all other ridges
- **Delta**: a Y-shaped ridge meeting
- **Pore, or independent ridge**: a ridge that commences, travels a short distance and then ends

Figure 2.21: Ridges and minutiae: (a) Ridges on real finger. (b) Minutiae identification on fingerprint

Even if all automated fingerprint identification systems use several biometric acquisition sensors (optical, capacitive, thermal or ultrasound sensors) and different analysis methods, the identification principle remains substantially the same. It consist in two main phase: extraction/encoding useful features, and compare features from two fingerprints.

The state of the art in fingerprint recognition contains two main categories.

The first category concerns the conventional approaches which simply compare the relative positions of minutiae. In [Jain 1997], the fist steps of processing consist in applying a directional filtering on fingerprint image, followed by a binarization and finally a by a thinning of ridges. The last step consists in determining minutiae positions on the image to quantify the similarity characteristics between two templates by “point pattern matching”. Maio et al. [Maio 1998] propose an alternative method to localize minutiae in a direct way by using neural networks.

The second category concerns more complex approaches which extract and use more information from the fingerprint, like local directions of some minutiae [Halici 1996] and [Capelli 1999], or the local frequential component of textures on images [Jain 1999]

2.3.1.3 Face Recognition

Face recognition has been the subject of many researches since many years, especially the two last decades. It has special attention from computer vision community due to its importance and effectiveness for people identification. The performances of face recognition systems have greatly increased since the first works in the 1960-1970s [Bledsoe 1964, Kelly 1971, Kanade 1977] and many new face recognition algorithms have been proposed since that time. These approaches can be categorized into three kinds according to the input information type: 2D-based approaches which use images and videos, 3D approaches based on 3D scanning, and finally hybrid approaches which combine both information. Unfortunately, these two last types of approaches, despite the highest performances that they can provide, have some drawbacks. The first one concerns high cost of 3D facial scanners. The second inconvenient is the high amount of information that have to be processed, which does not allow real-time processing. The last inconvenient concerns the unavailability of large datasets of 3D data for evaluation. For these reasons, most of studies are carried on 2D-based approaches, which are more practical.

Face recognition methods can be divided in two main categories according to the used information level: global (holistic) methods and local methods, based on models.

Global methods are based on statistical analysis and does not requires any face features localization or extraction (like eyes, nose, mouth, etc.). In this methods, the whole face image is processed as grid of pixels which is generally transformed to vector, more practical to process (See figure 2.22). This transformation is accompanied by a dimensional reduction to model face in a low dimensional sub-space, more significant and with faster processing. Global methods are divided in two main types of techniques which linear and non-linear techniques which determine.

Linear techniques project initial information to a lower sub-space linearly. The most known technique is the Principle Component Analysis (PCA). This technique was initially used for face representation in [Sirovich 1987, Kirby 1990] and was taken as ”Eigenfaces” technique in [Turk 1991]. Some other techniques based on linear decomposition have been used, like Linear Discriminant Analysis (LDA) [Belhumeur 1997] or Independent Component Analysis (ICA) [Bartlett 2002]. The linear techniques present the main issue of inability to conserve the geometric manifolds contained in original face images. This is due to the limitation to manage their non-linearity. To deal with this issue, linear methods have been extended to non-linear techniques based on “kernel” notion like **Kernel PCA** [Schölkopf 1998] and **Kernel LDA** [Mika 1999].

Global methods present the main advantage of being relatively faster, due to the medium complexity of required computations. However, they have are highly sensitive

to variations of illumination, pose, and facial expressions, which represent an important issue. This issue is due to the fact that processed information is a low level one, directly extracted from pixel values and not encoded to more robust and invariant features.

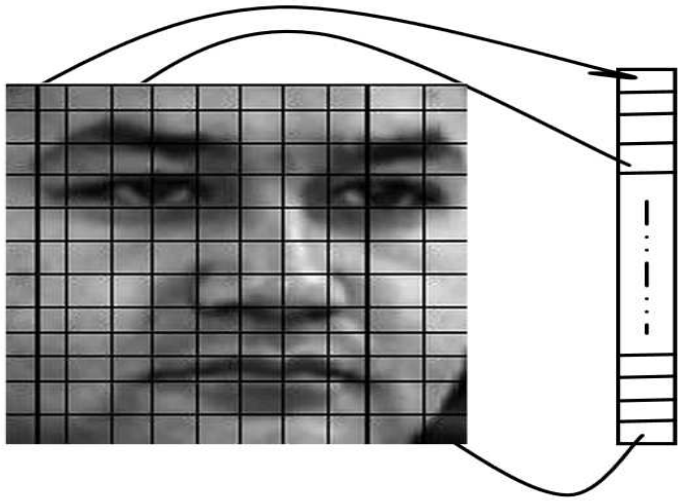


Figure 2.22: Global vector obtained from the whole image without any feature localization

Local methods are based on the a priori knowledge on the face morphology and use face features (See figure 2.23). Kanade [Kanade 1973] have proposed one of the first algorithms of this kind of approaches, by detecting some points and features on face image and by comparing them to points and features extracted from other face images. Other kinds of local approaches like Bayesian approaches [Liu 1998], Support Vector Machines (SVM) [Guo 2000], Active Appearance Models (AAM) [Cootes 2001] or Local Binary Pattern method (LBP) [Ahonen 2004] have been proposed. All these local methods have the advantage to model easily pose, illumination and face expression variations. Nevertheless, this kind of methods are more complex and requires more processing time in comparison with global methods.

Some works are based on a combination between global and local methods, providing hybrid method. Local Feature Analysis (LFA) [Penev 1996] and extracted features by oriented Gabor wavelets, like Elastic Bunch Graph Matching (EBGM) [Wiskott 1997] are some of hybrid methods examples. More recently, Log Gabor PCA (LG-PCA) [Perlibakas 2005] has been presented.

There are other approaches, based on neural networks [Lin 1997] or on Hidden Markov Models [Nefian 1998], but these approaches present important issues when the number of individuals increases, due to the complexity of computation. In addition to this, these two approaches requires many images for each face to train systems and to

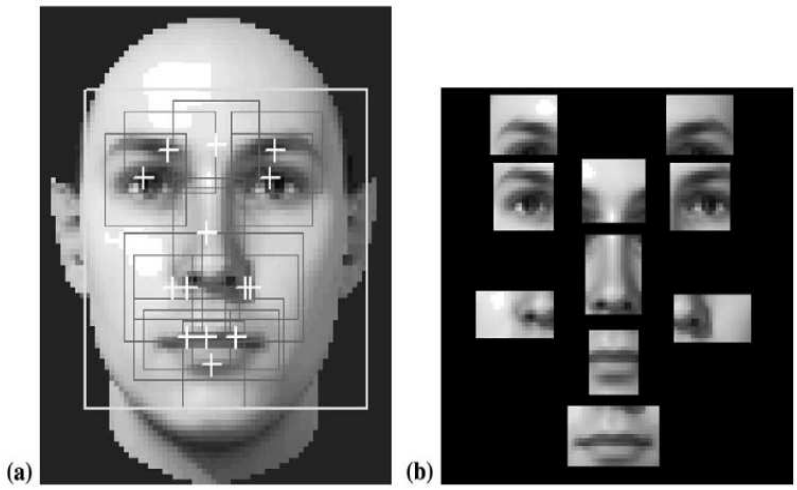


Figure 2.23: [Heisele 2003] face detection and recognition approaches: (a) The 14 components of the proposed face detector. The white crosses denote the center of each component. (b) The 10 used components for face recognition

configure parameters optimally.

2.3.1.4 Gait Recognition

Gait recognition addresses the problem of human identification by characterizing and discriminating the way they walk. Each person seems to have a distinctive way of walking. The differences in gait can be observed from many walking parameters, like gait cycle information (step frequencies and magnitudes) (See figure 2.24), person's pelvis/centroid vertical oscillations range, the maximum height of the foot when it leave the floor, etc.

One of the first attempts of automated gait analysis was performed in 1994 by Niyogi and Adelson [Niyogi 1993]. Individual gaits is recognized by applying standard pattern recognition techniques to the contour of individual, extracted by snakes approach. Many approaches have been published later. Lee et al. [Lee 2003b] perform a gait representation by a simple localization of image features such as moments extracted from orthogonal view video silhouettes of human walking motion. A suite of time-integration methods, spanning a range of coarseness of time aggregation and modelling of feature distributions, are applied to these image features to create a suite of gait sequence representations. Yoo et al. [Yoo 2002] generate gait signatures by extracted kinematic features in order to recognize people. He also propose a new method for extracting the body points by topological analysis and linear regression guided by anatomical knowl-

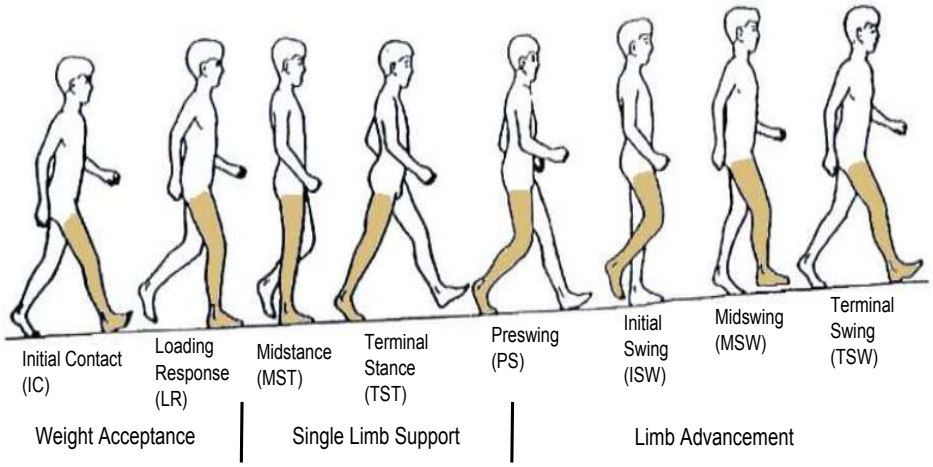


Figure 2.24: A complete gait cycle with its three tasks and eight phases displayed. (Source [Søndrål 2005])

edge.

In their approach, Canado et al.[Cunado 1997] consider legs only. The extract lines which represent legs on images using the Hough transform. The change in inclination of these lines follows simple harmonic motion; this motion is used as the gait biometric. The method of least squares is used to smooth the data and to infill for missing points. Then, Fourier transform analysis is used to reveal the frequency components of the change in inclination of the legs. The transform data is then classified using the k-nearest neighbour rule.

Kale et al. [Kale 2004] use two different image features: the width of the outer contour of the binarized silhouette of the walking person and the entire binary silhouette itself. From these two features, characterisation is performed following two different methods. In the first method, the high-dimensional image feature is transformed to a lower dimensional space. In second method, they work with the feature vector directly by training a Hidden Markov Model.

Wang et al. [Wang 2003a] use statistical shape analysis. Moving silhouettes of walking figures are extracted using a background subtraction algorithm. Temporal changes of the detected silhouettes are then represented as an associated sequence of complex vector configurations in a common coordinate frame, and are further analysed using the Procrustes shape analysis method to obtain mean shape as gait signature. Supervised pattern classification techniques based on the full Procrustes distance measure are finally used for recognition.

[Liu 2004] simply align and average the silhouettes over one gait cycle (See figure 2.25). The recognition is then performed using the Euclidean distance between these averaged silhouette representations.

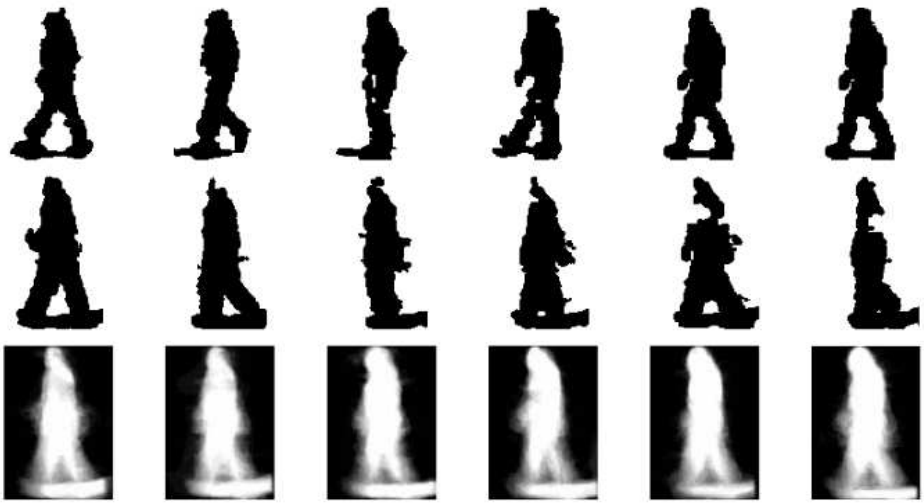


Figure 2.25: [Liu 2004] average silhouette computation: First and second rows show samples of the binary silhouettes over one gait cycle, for two subjects, respectively. The third row shows the averaged silhouettes for the subject in the second row; each averaged over a different gait cycle.

2.3.2 Appearance-based approaches

Despite their effectiveness, the hard constraints of biometric approaches, presented in the previous section, make this kind of approaches inadequate for large wide video-surveillance purpose. The alternative consists in using global appearance information, which seems to be "easier" to extract in term of constraints. Nevertheless, despite the availability of this information, a major problem remains: how to use the global appearance information to characterize an individual in a robust and reliable way? As people re-identification concerns a large set of individuals acquired from different cameras, under several conditions, it is necessary to provide a distinctive and invariant to camera changes signature. This visual signature has to be invariant to some external variations, like changes in illumination, person pose and orientation and camera point of view. It has to be robust against internal camera parameters also. The sensitivity of different camera sensors and the color acquisition vary from a camera to another one, providing different images of the same person in terms of color (See figure 2.26). This last issue has been handled in two different ways in appearance-based re-identification methods.

Many approaches use some color normalization methods without any a priori information about camera parameters, while other methods are based on multi-camera color calibration, using some colorimetric transfer functions.

Figure 2.26: Difference in color rendering between cameras: The same person observed by two different cameras. (Images from iLids multi-camera tracking dataset).

In appearance-based re-identification approaches, a visual signature of a given person consists in a set of extracted features from a single or multiple images of him. The choice of these features and the manner in which they are combined to obtain a discriminative visual signature is the main challenge. The visual signature of a given person has to be as restrictive as possible to maximize inter-class distance between different person signatures, and at the same time, sufficiently permissive to minimize intra-class distance between possible appearances of the same person observed under different conditions and by different cameras. Many approaches with different levels of complexity have been proposed to address this challenge.

The simplest approaches are based on statistical computations on the whole person images, using low level information and thereby, avoiding spatial information, applying a histogram representation. More complex techniques uses more sophisticated features, taking into account spatial information which is very useful to solve matching problem. This spatial information can be classified into two categories: As a result of feature extraction techniques or as an initialization for them. In the first case, feature extraction does not use any a priori information concerning the localization of the wanted features. The spatial information is obtained once the features are extracted, for instance the coordinates of points of interest, the coordinates of the centroid and the area of a stable color region, etc. In the second case, features are extracted in predefined regions. They can be the result of uniform subdivision of the person image (horizontal strips in [Bird 2005] [Truong Cong 2010] or grid [Bak 2011]) or a specific region delimitation like human body parts [Bak 2010] or symmetry/asymmetry body subdivision [Farenzena 2010].

The more sophisticated approaches can be categorized in two main groups accord-

ing to the way on which they focus to perform the people re-identification. The first category concerns the Feature Oriented (FO) approaches, which focus on designing an invariant feature, which should handle viewpoint and camera changes [Bazzani 2010] [Farenzena 2010]. The second category contains approaches which concentrate on learning aspects or on feature modelling. Learning approaches use training data of different individuals to select the best features and to find the best way to combine them. These approaches focus either on metric learning for matching appearances regardless of the representation choice [Dikmen 2011], or on discriminative methods which enhance discriminative features of a specific individual [Schwartz 2009].

Another classification of the existing approaches can be performed, based on the number of used images per person. According to this creteria, two families of approaches for appearance-base people re-identification exist: Single-shot approaches which extract information from a single image of a person and Multiple-shot approaches which use information of multiple images of a person to encode possible variations and learn a reliable representation of the person.

The two classification point of views intersect as much as some feature oriented approaches use single images per person while other feature oriented approaches are based on multiple images per person. Similarly, some learning approaches are performed using single images per person while other approaches require multiple images per person. We take the second classification point of view as a basis for the presentation of the state of the art, as long as it relies on a primary parameter: the number of used images per person. Nevertheless, before exposing single-shot and multiple-shot state of the art, a brief presentation of existing works for colorimetric transfer function is presented first. This task is important as long as it can allows to deal with difference in color acquisition between different cameras.

2.3.2.1 Colorimetric Transfer Function

As mentioned before, the acquired images of the same person by two cameras can present a significant difference in rendered colors (see figure 2.26). It may be due to sensor sensitivity difference or to external reasons, like difference in illumination or camera orientation. This issue becomes critical as soon as a re-identification process requires to compute a color-based people signatures.

To handle this issue, most of appearance-based people re-identification approaches integrate a color normalization step in their process, generally before the extraction of any feature. However, many researches have been done to provide a “link” between different color rendering. This task is called “Colorimetric calibration” between different cameras.

Porikli [Porikli 2003] proposes an initial method for colorimetric calibration between different cameras, called “Brightness Transfer Function”. He proposes a distance metric and a model function to evaluate the inter-camera radiometric properties. Instead of depending on the shape assumptions of brightness transfer function to find separate radiometric responses, he derives a non-parametric function to model color distortion for pair-wise camera combinations. This method is based on correlation matrix analysis. The correlation matrix is computed from three 1-D color histograms, and the model function is obtained from a minimum cost path traced within the matrix. The model function enables accurate compensation of color mismatches, which cannot be done with conventional distance metrics. An illustration is given in figure 2.27

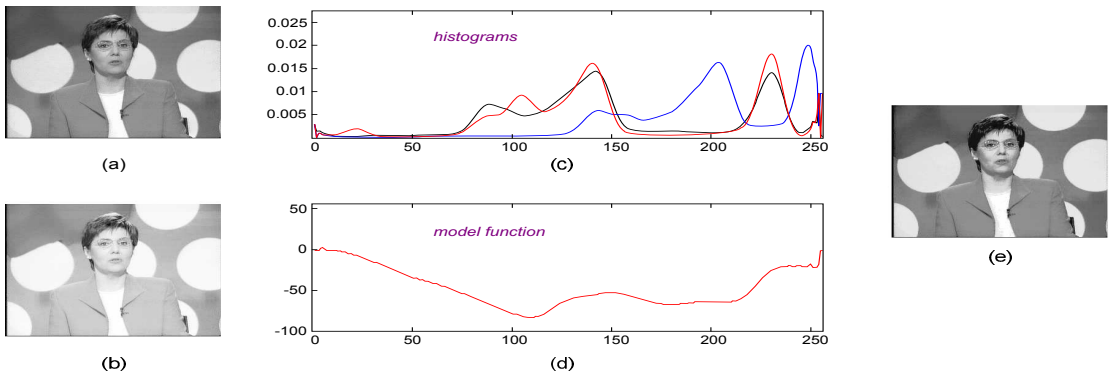


Figure 2.27: [Porikli 2003] example of Brightness Transfer Function application on one channel image: (a) Reference image. (b) over-exposed image. (c) Intensity histograms of the input image (shown as black), of the overexposed image (blue), and of the compensated image (red). (d) The model function that maps the over-exposed image to the original (red). (d) The compensated image.

[Gilbert 2006] approach is based on [Porikli 2003], but integrates an on-line learning phase to update illumination variations between cameras, extending the use of correlation matrices. However, this approach is greatly sensitive to the initialisation phase of the Transfer Function and requires a large training dataset (more than 5000 trajectories).

[Javed 2005] and later [Javed 2008] propose another extension of [Porikli 2003] approach. In this approach, a Brightness Transfer Function is estimated for each pair of images of the same person acquired by two different cameras under different illumination conditions. It obtains as many BTFs as image pairs number. The final Brightness Transfer Function, which provides the best representation of the color changes between cameras is computed by applying a Principle Component analysis (PCA) on all the individual Brightness Transfer Functions.

These previous approaches assume that the considered objects are observed with relatively same point of view by the different cameras. Prosser et al. [Prosser 2008] address a more general case where objects are not observed with the same point of view. In this case, the observed proportions of object colors may vary greatly from a camera to another one. [Prosser 2008] propose a Cumulative Brightness Transfer Function (CBTF). It cumulates many images of the same person, observed by the same camera, in the same histogram, before applying Porikli's method on the cumulated histograms, obtaining the Cumulative Brightness Transfer Function.

Colorimetric calibration presents two main issues for our addressed problematic. First, colorimetric transfer functions are generally not bijective depending on the acquisition conditions (camera orientation, lighting conditions, etc.). Some large ranges of colors from a given camera may have unique color (in case of discretization) or a smaller range of colors as correspondence. This phenomenon is most observable for very low/high saturation values. This decrease the efficiency of colorimetric calibration. The second issue is more important for our work since it is related to the difficulty to apply this kind of approaches in large scale video-surveillance systems. It requires to annotate a sufficient pairs of persons observed by each pair of deployed cameras. This is not conceivable as a generic method since these transfer functions are not related to the internal cameras only, but also to external conditions (light conditions, orientations, etc.).

In the following paragraphs, the main approaches of the state of the art in appearance-based re-identification are presented. These approaches are divided in two categories: Single-shot approaches and Multiple-shot approaches.

2.3.2.2 Single-shot Approaches

In the single-shot methods, the re-identification is performed using a single image for each person. As mentioned before, two categories of approaches can be distinguishes: feature-oriented approaches and learning approaches.

- **Feature-oriented approaches**

  Feature-oriented (FO) approaches focus on finding the best feature representation to be invariant to the possible variation.

  In [Park 2006], the proposed approach, called Visual Search Engine (ViSE), is based first on a segmentation of the person's body in three parts: the head, the

torso, and the legs. This segmentation aims at keeping spatial correlation of feature distributions and is based on approximative dimensions of each body part with respect to the whole height (the two separation lines are at 1/5th and 3/5th of the person height). The head part is ignored due to the low discriminative power of its features when no biometric techniques are used. Each shirt and pants region (corresponding respectively to the torso and the legs) is characterized with a 10 bins color histogram in HSV space. These 10 bins correspond to most distinguishable colors by human and are: red, brown, yellow, green, blue, violet, pink, white, black and gray. The final color for each shirt and pants region is decided as the bin with the largest count. A schematic diagram of the proposed ViSE is presented in figure 2.28

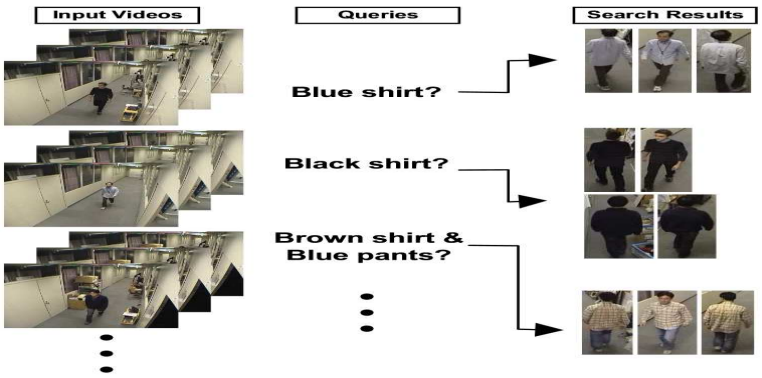


Figure 2.28: Schematic Diagram of ViSE [Park 2006]

In [Gallagher 2008], a collaboration between clothing segmentation and characterization in one side, and facial features in the other is performed to recognize individuals. The face detection is performed using [Viola 2003] and the clothing segmentation is obtained using graph cuts and clothing mask. Each face is characterized by a 37-dimensional vector, obtained by projecting the face image onto a set of Fisherfaces [Belhumeur 1997]. Clothing regions are represented by 5-dimensional feature vectors corresponding to a linear transformation of the three RGB color values and the responses to a horizontal and vertical edge detector. The final appearance is represented by the set of histograms over each of the 5 features on the segmented image. Examples of the clothing segmentation is presented in figure 2.29.

In [Cai 2008], the proposed approach uses color patches to represent the person on interest. These patches are localized along edges, which are extracted using Canny edge detection algorithm [Canny 1986]. Each region is represented by the

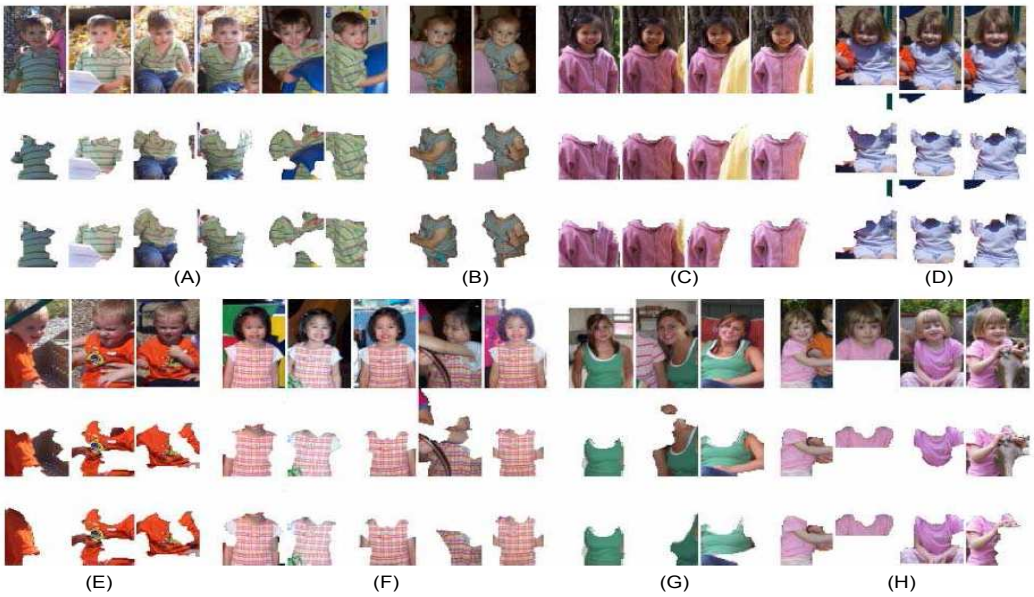


Figure 2.29: Clothing segmentation using graph cuts [Gallagher 2008]: For each group of person images, the top row shows the resized person images, the middle row shows the result of applying graph cuts to segment clothing on each person image individually, and the bottom row shows the result of segmenting the clothing using the entire group of images.

dominant color and its frequency in this region. The top of the head is taken as reference point to encode the spacial correlation between edge points. The spatial information of each candidate point is encoded using the distance D between it and the head point in addition to the angle θ (See figure 2.30). To be invariant to scale variations, the distance between the head point and any edge point is normalized by the height of the silhouette. This spatial correlation is used in addition to color characterisation of patches to perform matching.

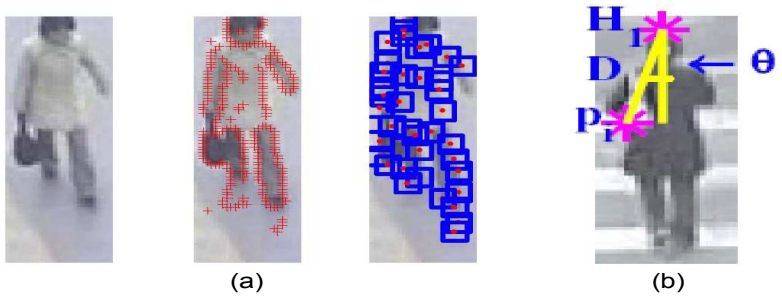


Figure 2.30: Human appearance [Cai 2008]: (a) From left: original image, results of Canny edge detection, region signatures on the edges; (b) geometric constraints.

[Yu 2007] present an appearance model for human re-identification. The appearance model is constructed by kernel density estimation. To incorporate structural information and to achieve invariance to motion and pose, besides color features, an additional feature of path-length of the pixels inside the silhouette of a person is used, by taking the top the head as a reference point. To achieve illumination invariance, two types of illumination insensitive color features are tested: brightness color feature and RGB rank feature. The similarity between a test image and an appearance model is measured by the information gain or Kullback-Leibler distance [Kullback 1968]. To thoroughly represent the information contained in a video sequence with as little data as possible, a key frame selection and matching scheme is proposed.

In [Kang 2004], they model the object of interest for reacquisition purpose using stochastic models. The appearance of the object is described by multiple models representing spatial distributions of objects' colors and edges. It is performed using the smallest circle containing the object blob. This circle is uniformly sampled into a set of control points, from which a set of concentric circles of various radii are used for defining bins of the appearance model. Inside each bin, a Gaussian color model is computed for modelling the color properties of the overlapping pixels of the detected blob (see figure 2.31).

[Wang 2007] introduce the concept of shape and appearance context (See figure 2.32). A pedestrian image is segmented into regions and their color spatial information is registered into a co-occurrence matrix. A region appearance is represented by histogram of oriented gradients (HOG) in the Log-RGB color space [Funt 2002]. Parts identification is done by modified shape context algorithm [Belongie 2002], which uses a shape dictionary learnt a priori. The context of the appearance and shape is handled by using occurrence/co-occurrence function which describes probability distributions and their correlations over the image region.

[Bak 2010] approach is base on spatial covariance regions extracted from human body parts. A human body part detector, based on Histogram of Oriented Gradient technique (HOG) [Corvee 2009] is trained and applied to detect 5 body parts: the top, the torso, legs, the left arm and the right arm. To handle color dissimilarities caused by camera and illumination differences, a color normalization technique called histogram equalization [Hordley 2005] is applied. After that, the covariance regions of body parts are computed on normalized images to generate a human signature. The dissimilarities between these regions corresponding to different

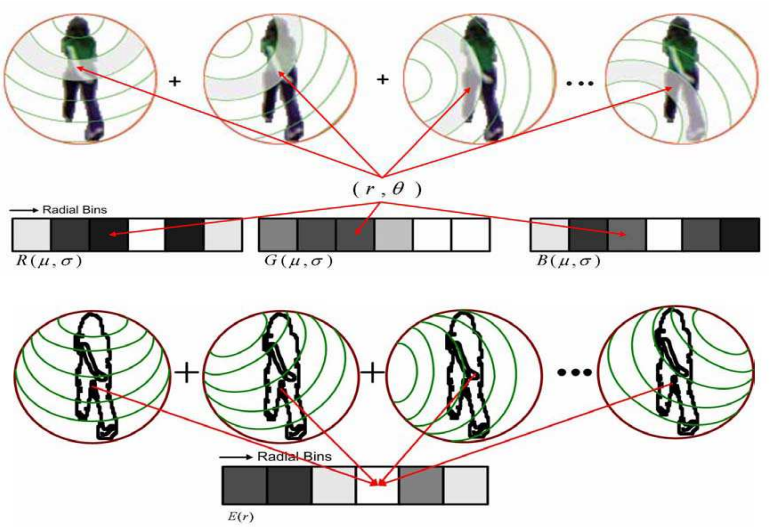


Figure 2.31: Computation of the color and shape based appearance model of detected moving blobs [Kang 2004]

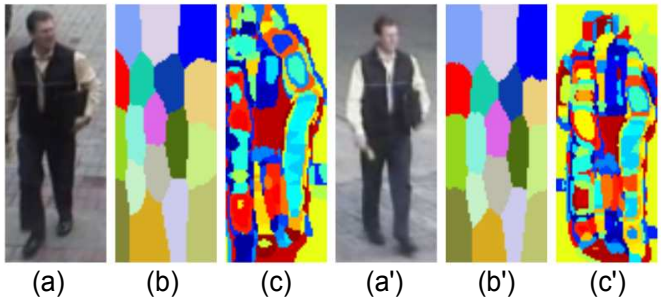


Figure 2.32: [Wang 2007] Shape and appearance labelled images. (a) and (a') are two images of the same person, acquired by two different cameras at different locations. (b) and (b') are their corresponding shape labelled images respectively. (c) and (c') are their appearance labelled images respectively.

images are combined using an idea derived from the spatial pyramid match kernels [Grauman 2005] (See figure 2.33).

■ **Learning approaches**

Learning approaches require training data to find the best way to perform appearance matching. We can distinguish two types of learning methods, according to the aspect on which they focus: Some approaches concentrate on metric learning (ML) for matching regardless of the person representation choice, while other approaches focus on the way to enhance discriminative features of an individual and are called discriminative methods (DM).

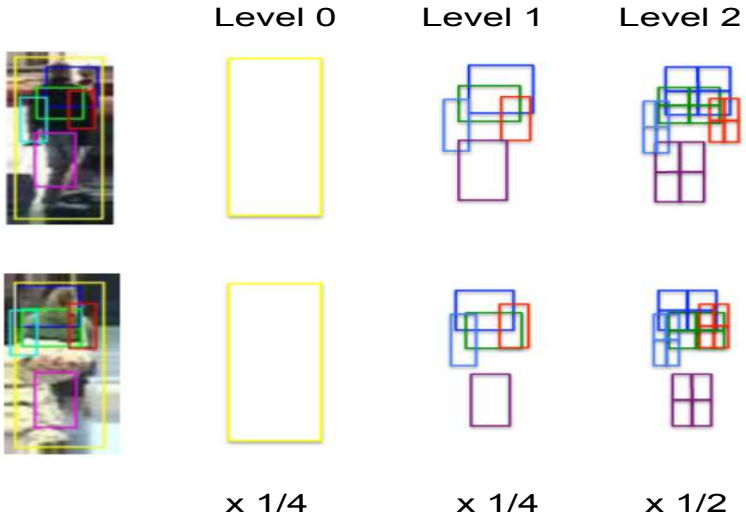


Figure 2.33: Example of constructing a three-level pyramid. The left column represents two person images with the detected body parts. Level 0 corresponds to the full body part. Level 1 and level 2 correspond to the rest of detected body parts and grids inside body parts, respectively [Bak 2010]

From the recent ***metric learning (ML)*** approaches, we can cite [Dikmen 2011, Hirzer 2012, Ijiri 2012] works. [Dikmen 2011] use a metric learning framework to obtain a robust metric for large margin nearest neighbor classification with rejection (i.e., classifier will return no matches if all neighbours are beyond a certain distance). The rejection condition necessitates the use of a uniform threshold for a maximum allowed distance for deeming a pair of images as a match. In order to handle the rejection case, they propose a novel cost function called Large Margin Nearest Neighbour with Rejection (LMNN-R), similar to the Large Margin Nearest Neighbour (LMNN).

[Hirzer 2012] address the problem of cameras point of view and property differences by learning the transition from one camera to the other. The human representation is performed using HSV and Lab color channels for color information and LPB for texture information. The human image is divided on a grid of 8x16 rectangular regions, using an overlap of 50 % between regions in both directions. In each rectangular patch, the mean values per color channel is calculated and discretized to the range 0 to 40. Additionally, a histogram of LBP codes is generated from a gray value representation of the patch. These values are then put together to form a feature vector. The vectors from all regions are concatenated to build a representation for the whole image. A PCA is applied for dimension-

ality reduction. The proposed classification method is built on the ideas of Large Margin Nearest Neighbour classification solution. This is realized by learning a Mahalanobis metric using pairs of labelled samples from different cameras.

[Ijiri 2012] proposes an other metric learning based approach to handle the same issue. To learn the optimal metric, human body images are divided into several vertically segmented regions following [Bird 2005] approach. The color histogram of each body region is computed in HSV color space and the entire person representation is obtained by concatenating all the HSV part histograms. The corresponding label is assigned to each person histogram representation, providing a training dataset. The optimal metric learning is based on Large Margin Component Analysis (LMCA) [Torresani 2006]. A non-linear projection is used to project the input histograms onto a higher dimensional space. For non-linear projection, several types of kernel functions were investigated (X^2 kernel, Bhattacharyya kernel, Jeffrey divergence kernel, Jensen-Shannon kernel). This approach is illustrated in figure 2.34.

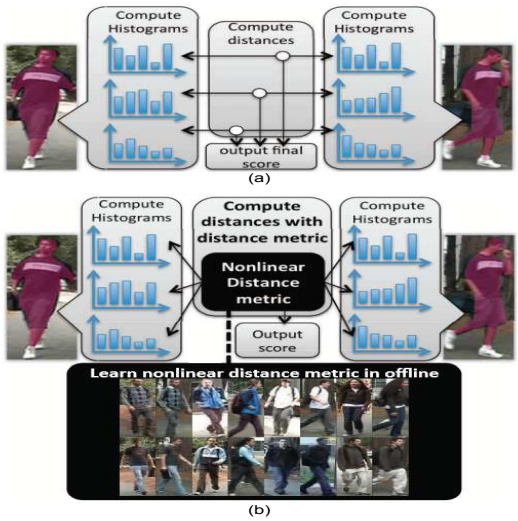


Figure 2.34: [Ijiri 2012] metric learning for re-identification: (a) conventional and (b) proposed color matching schemes

Metric learning approaches are usually performed off-line. The positive data consists in pairs of images of the same person acquired by different cameras, and negative data consists in pairs of images of different person acquired by different cameras.

Concerning ***Discriminative methods (DM)***, we can cite [Lin 2008, Schwartz 2009] approaches. In [Lin 2008], the person appearance is represented with a 4-dimensional

vectors, containing 3 color components and the height coordinate. Features are aggregated using the probability density function (PDF). The distance between two appearances is established using pairwise dissimilarity profiles which are learned beforehand. The nearest neighbour classification is adapted to perform re-identification.

In [Schwartz 2009], a rich set of feature descriptors based on color, textures and edges is used. Features are extracted from overlapping blocks constructing a high-dimensional feature vector. The high-dimensional signature is transformed into a low-dimensional discriminant latent space using a statistical tool called Partial Least Squares (PLS) in one-against-all scheme (the discriminative appearance of person is learned using information about the appearances of other persons). For the one-against-all scheme, PLS gives higher weights to features located in regions containing discriminative characteristics.

Opposite to Metric Learning (ML) approaches, the discriminative approaches learning is usually performed on-line.

2.3.2.3 Multiple-shot Approaches

In the multiple-shot methods, the re-identification is performed using multiple images of the same person. This allows to better take into account appearance variations of the same person and thereby, provides a more informative signature. As single-shot approaches, multiple-shot approaches can also be divided into two classes: feature-oriented approaches and learning approaches.

- **Feature-oriented approaches**

  In the multiple-shot case, the availability of more than one image per person allows the use of powerful mathematical tools, such as clustering, Principal Component Analysis and many other ones, to extract most reliable features.

  In [Gheissari 2006], the human representation is based on edge extraction. A spatio-temporal graph which uses multiple images is proposed to group spatio-temporally similar regions. Temporally unstable edges are rejected using spatio-temporal segmentation (see figure 2.35 (a)). Only edges which are interior to the foreground are considered. Then, a triangulated person model is used to handle a correspondence between different body parts. The person model is represented by a decomposable triangulated graph as a method for model fitting to people (see figure2.35 (b)). A dynamic programming algorithm is used to fit the model to the image of the person. Image regions are compared using color and structural

information. The color information is represented by normalized histograms based on hue and saturation. The structural appearance is captured using edge detector.

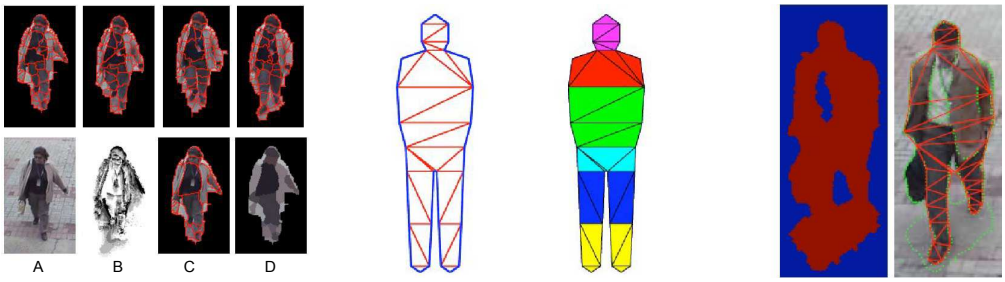


Figure 2.35: Spatio-temporal appearance [Gheissari 2006]. (a) Upper row: segmentation for single frames. Lower row: A) original image, B) frequency image, C) final segmentation after graph partitioning, D) median image for final segmentation; (b) left: an example of a decomposable triangulated graph used as a person model. The blue edges correspond to the boundary of the person while the red edges are interior edges. Right: partitioning of the person into body parts. (c) Left: foreground mask. Right: fitting results.

In [Hamdoun 2008], the person signature consists in an accumulation of SURF points of interest [Bay 2008] extracted from multiple images of each person. These cumulated points of interest are stored in KD-tree to speed-up the query processing time (see figure 2.36 (a)). The association of the models is obtained by a voting approach: every interest point extracted from the query is compared to all models points stored in the KD-tree, and a vote is added for each model containing the nearest descriptor. Finally the re-identification is performed with the highest vote for the model (see figure 2.36 (b)).

In [Huang 2009], the person image is divided into three parts: the head, the torso, and the legs. This segmentation is performed using approximative ratios of 1/5, 2/5 and 2/5 (see figure 2.37 (c)). The head is ignored due to the low amount of extractable information without using biometric techniques. From the two remaining parts, a tree structure containing medians of RGB colors is extracted. The median value of a given node is used to separate its corresponding histogram and thereby creating the two child histograms. The final appearance feature is a vector obtained by merging median vectors. Finally, a Bayesian-based tracker combines a set of features using a multivariate normal distribution.

In [Farenzena 2010], proposes another method to divide person body image into three main parts which are the head, the torso and the legs. The head is ignored in this approach too. This segmentation is based on maximizing the Euclidean distance between the colors of two adjacent parts (head/torso and torso/legs) while

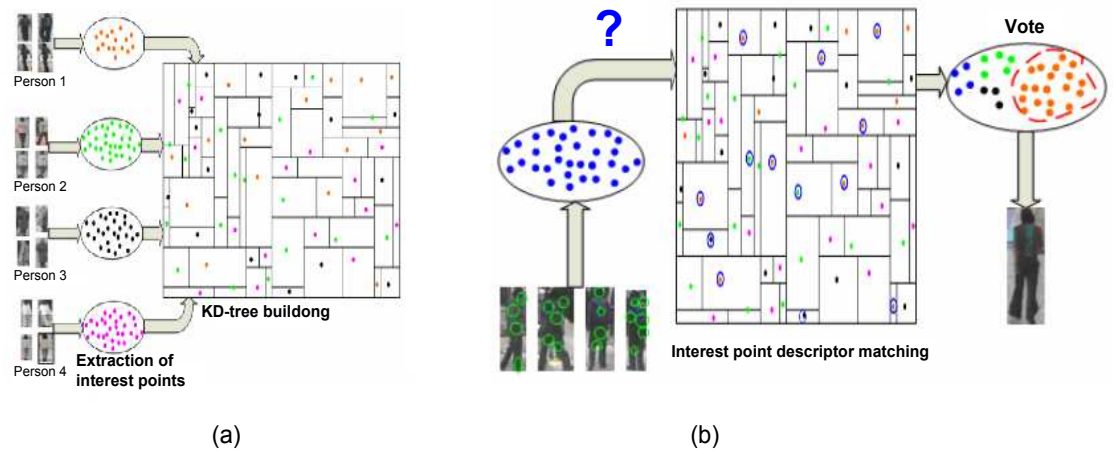


Figure 2.36: [Hamdoun 2008] people re-identification process: (a) Schematic view of model building. For each tracked person, interest points are collected from multiple images, and the personś signature is the union of all these key-points stored in a global KD-tree. (b) Schematic view of re-identification of a query. Interest points are collected similarly multiple images images of the person to be re-identified, and a vote is made according to the respective identifications associated to all matching key-points

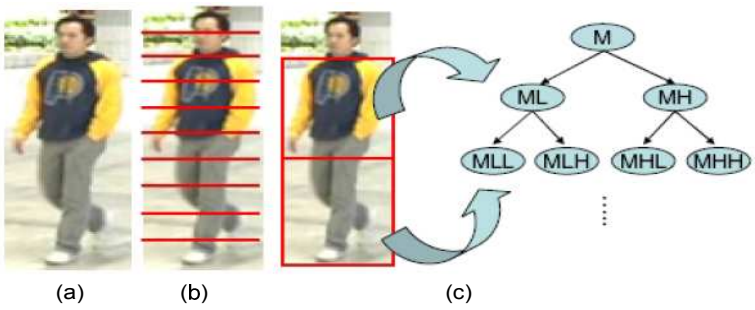


Figure 2.37: Segmented parts of the human body: (a) original image; (b) 10 horizontal strips [Bird 2005]; (c) The tree structure extracted from upper and lower body In [Huang 2009].

minimizing normalized areas difference between these two (the normalization is obtained by the ratio between foreground pixel of the part and its whole bounding box rectangular area, using a mask). Another segmentation, using symmetry axis, is performed on torso and legs to ensure robustness against people rotations (see figure 2.38(b)). The appearance of each body region is represented by the combination of three features: chromatic content (HSV histogram) (see figure 2.38(c)); maximally stable color regions (MSCR) [Forssén 2007] (see figure 2.38(d)) and (3) recurrent highly structured patches (RHSP)(see figure 2.38(e)). The final matching score between two extracted signatures is the weighted sum

of distances between every couple of similar features, belonging to the two signatures.

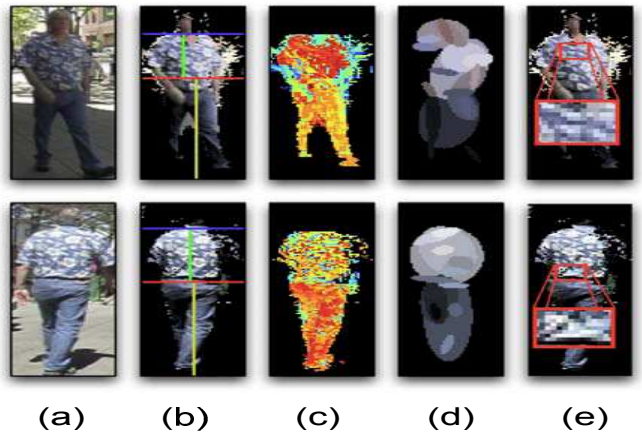


Figure 2.38: Sketch of SDALF approach [Farenzena 2010]: (a) two instances of the same person; (b) x- and y-axes of asymmetry and symmetry, respectively; (c) weighted histogram back-projection (brighter pixels mean a more important color), (d) Maximally Stable Color Regions; (e) Recurrent Highly Structured Patches.

■ **Learning approaches**

In multiple-shot approaches, the different approaches often use dimensionality reduction methods, classification using SVM or Boosting and Fisher discriminants to discriminate between different individuals.

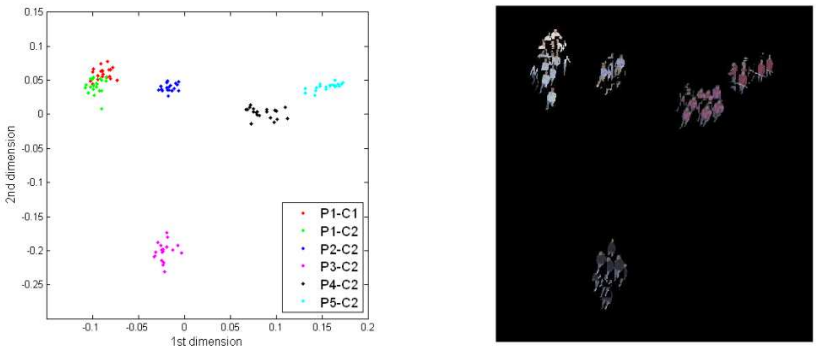


Figure 2.39: Example obtained by the graph-based approach for nonlinear dimensionality reduction. Multiple images of 5 different persons, acquired by two cameras are used.[Truong Cong 2009]

In [Nakajima 2003], images of full-body persons are represented by color-based and shape-based features. Recognition is carried out through combinations of

Support Vector Machine (SVM) classifiers. Different types of multi-class strategies based on SVMs are explored and compared to k-Nearest Neighbours classifiers. The best performance is obtained using two dimensional normalized color histograms, where d1 = R=(R + G + B) and d2 = G=(R + G + B) build two dimensional space (every dimension is represented by 32 bins)

In [Truong Cong 2009], the approach uses different color-based features, combined with several illuminant invariant normalizations in order to characterize the silhouettes in static frames. A color-based feature vector is extracted from each image of each person. The hight dimensional extracted vector features are projected to a lower dimensional space using a proposed graph-based approach which is capable of learning the global structure of the manifold and preserving the properties of the original data in a lower dimensional. Each person is represented by a set of points in the lower dimension space (see figure 2.39), and the centroid of these points is taken as the reference point for signature comparison. The dissimilarity measure between two signature is computed using the distance between the two corresponding centroids.

In [Truong Cong 2010], the authors propose two improvements to their previous approach [Truong Cong 2009]. They use the same graph-based approach for dimensionality reduction as in [Truong Cong 2009], but this time, it has been applied on a new proposed descriptor called the color-position histogram. The human body is vertically divided into n equal parts and the mean color is computed to characterize each part(see figure 2.40). This approach allow to use spatial information of color ant thereby, it provides better results than a classical color histogram. The second improvement concerns the dissimilarity measure between two signatures. Instead of using the distance between centroids of points groups, the authors use the optimal margin and the miss-classification error obtained by SVM to compute distance between signatures which can not be separated by a linear model.

In [Bak 2011], a new appearance model based on Mean Riemannian Covariance (MRC) is proposed. Person images are obtained by detecting and tracking them using an HOG-based algorithm [Corvee 2009]. Once multiple images of each person are extracted, a color normalization technique called histogram equalization [Hordley 2005] is applied on them. This step aims at minimizing illumination and color acquisition variations between cameras effect (see figure 2.26). All the images are then resized to the same dimensions. Two set of patches are extracted from each image, using two specific patch sizes and using an overlapping shift. A

Figure 2.40: Color-position histogram [Truong Cong 2010]: (a) original image; (b) localization of the silhouette; (c) color distribution in the silhouette.

mean covariance is computed for each patch location, using all patches at the same location, upon all the images of each person (see figure 2.41). A person is represented in a first time by a set of mean covariance regions of all its patches. The final person signature is obtained by selecting the most significant patch and removing most variable ones, assuming that these last patches are the noisiest ones. Authors propose two ways to perform this selection. First, a reliability measure on patches is introduced. For each region, this reliability measure is computed using the standard deviation which is associated to the mean covariance. Only patches with a high reliability measure are used for person signature (see figure 2.42(b)). The second way for significant patches selection is performed using a boosting scheme. Finally, a new similarity measure between signatures using Riemannian manifold theory is proposed. It enables to hold discriminative power coming from the relative position of MRC patches.

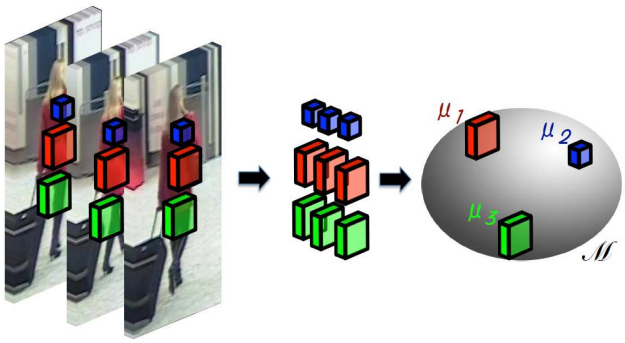
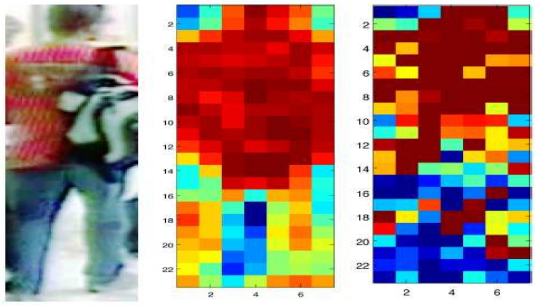


Figure 2.41: [Bak 2011] computation of three MRC patches. Covariances gathered from tracking results are used to compute the mean covariance using Riemannian manifold space (depicted with the surface of the sphere). The mean covariance forms MRC patch.



(a) (b) (c)

Figure 2.42: Illustration of patch significance: (a) one of many frames obtained during tracking; (b) reliability map obtained by the first method; (c) confidence map obtained by boosting. Colours correspond to significance of patches (for clarity only $\frac{W}{4}$x$\frac{W}{4}$ patches, shifted by $\frac{W}{8}$ pixels are illustrated, red indicates the highest significance, blue the lowest).

2.3.2.4 Context-Aware Approaches: Group Context / Space-time Approaches

All the previous cited approaches, both in single-shot and multiple-shot categories focus exclusively on the visual information contained in images of persons of interest only. Some other works, called Context-Aware approaches, use additional information to increase the discriminative power of their appearance-based approaches. This information can be classified into two groups: Group context information and Space-time information.

Group context information still uses visual information extracted from images, but does not focus on the person of interest silhouette only. It takes into account surrounding people also. This kind of approach is relatively recent in comparison with all previous cited techniques and thereby, very few works have been published. However, it will certainly attract more attention in the future, due to the improvements that they can provide, according to published works results, which are still medium, but which open many perspectives (keeping the main idea of using group information, but proposing more sophisticated techniques to deal with non-rigidity of people in a group, partial occlusion depending on the camera position, etc.).

In [Zheng 2009], a novel people group representation and a group matching algorithm are proposed to reduce ambiguity in person identification. Assuming that in a crowded public space, people often walk in groups, either with people they know or strangers, visual information coming from the surrounding is also used. The appearance of the group is represented by visual words. First, each pixel is represented by a feature vector consisting in a concatenation of SIFT [Lowe 2004] features for each RGB channel of this pixel. Then, a code book of n visual words is built by quantizing the previous fea-

ture vectors into n clusters by k-means. Finally, the group is represented by combining two descriptors: *a center rectangular ring ratio-occurrence descriptor* which aims at describing the ratio information of visual words within and between different rectangular ring regions, and a *block based ratio-occurrence descriptor* to explore more specific local spatial information between visual words that could be stable.

In [Cai 2010], covariance descriptor [Tuzel 2006] are used to encode group context information to improves person re-identification performances. The appearance of a person as well as the appearance of a group have been represented using covariance descriptor.

Space-time information is another useful information which allows better people re-identification performances. It can be used either as a filtering method, to reduce the space of possible candidates for a given re-identification request before applying appearance-based matching, and this by removing all candidates which do not respect spatio-temporal coherency (A person cannot be at two different locations at the same time. He/She cannot travel a given distance in a time which corresponds to incoherent velocity, etc.), or as an identification information (especially when camera fields of view overlap) by validating the same location of the same person at the same moment in the different camera images. The spatio-temporal information may be provided as an input information (floor plans with distances and possible paths provided to the system which manage the camera network), or learned automatically by the re-identification methods.

In [Javed 2003], a novel approach for establishing object correspondence across non-overlapping cameras is proposed. The multi-camera tracking algorithm exploits the redundancy in paths that people and cars tend to follow (e.g. roads, walk-ways or corridors), by using motion trends and appearance of objects, to establish correspondence. The proposed method does not require any inter-camera calibration, instead the system learns the camera topology and path probabilities of objects using Parzen windows, during a training phase. Once the training is complete, correspondences are assigned using the maximum a posteriori (MAP) estimation framework. The learned parameters are updated with changing trajectory patterns.

[Iwama 2012] propose a combined approach using group and space-time information. The relationships between the people in an input sequence are modelled using a graphical model. The identity of each person is then propagated to their neighbours in the form of message passing in a graph via belief propagation, depending on each personś group affiliation information and their characteristics, such as spatial distance and velocity vector difference, so that the members of the same group with similar characteristics enhance each otherś identities as group members.

	feature oriented	learning
single-shot	[Cai 2008] [Kang 2004] [Park 2006] [Yu 2007] [Wang 2007] [Gallagher 2008] [Bak 2010]	[Lin 2008] [Schwartz 2009] [Dikmen 2011] [Hirzer 2012] [Ijiri 2012]
multiple-shot	[Gheissari 2006] [Hamdoun 2008] [Huang 2009] [Farenzena 2010]	[Nakajima 2003] [Truong Cong 2009] [Truong Cong 2010] [Bak 2011]

Table 2.2: Summary of cited appearance-based approaches for people re-identification.

context-aware	group context	[Zheng 2009] [Cai 2010]	[Iwama 2012]
	space-time	[Javed 2003]	

Table 2.3: Context-Aware approaches for people re-identification

2.3.3 Discussion

In this section, we have presented several existing approaches for people re-identification. These approaches are classified according to a first level criterion which is the nature of used information, and each criteria is categorized using many sub-criteria.

We have seen that the first family of approaches, belonging to biometrics methods (iris, finger print, face and gait recognitions), are the most powerful methods to perform the people re-identification task, due to the discriminative power of human biometric characteristics. At the same time, these methods are the most constrained ones also. They require many conditions which cannot be ensured most of the time in video surveillance systems. The most important constraints we can cite are the voluntary collaboration of individuals (the acquisition of iris and fingerprint images by specific sensors), the restricted situation in which these approaches can be performed (visible faces for face recognition, walking direction allowing the segmentation of gait, i.e. as far as possible from to the camera optical axis direction), high resolution images (to provide face features) and high frame-rate acquisition (to capture all the gait cycle parts).

The second family of approaches, related to appearance-based approaches, are less

constrained than biometric ones and thereby can be used on large wide surveillance camera systems. These approaches are separated into single-shot approaches which require only a single image of a person to model his/her appearance, and multiple-shot approaches which requires multiple-images. In both of these categorises, the appearance representation can be extracted using feature-oriented procedures or by learning approaches.

Another group of approaches, called context-aware approaches, integrate additional information coming from environment to achieve people re-identification. Group-context information is used to improve individual signature by integration his surrounding people. Spatio-temporal information is also used to increase the reliability of re-identification.

Multiple-image approaches seem to be the ones which provide better results, due to their capability to encode the possible variation of human appearance and to extract/learn the most significant appearance model. Nevertheless, they have three main issues:

- Their computation cost is relatively high. The complexity of the extracted features and the learning techniques require complex computations, slowing down greatly the re-identification process. On the other hand, many approaches require the availability of a large set of images of a given person person of interest and a large set of other people images before performing the learning, i.e. they cannot compute the visual signature of a given person progressively, starting with the first image of this person and updating the signature every time a new image of him/her or of other people is provided (acquired by mono-camera tracking algorithm) . These issues make the use of this kind of approaches hard or impossible for live processing purpose. They are more suitable for off-line tasks like a posteriori people searching, where the processing time is less important and all people images can be extracted in a first processing pass, before performing re-identification in a second pass.

- They are sensitive to the precision of extracted images of the same person. Images of a given person are extracted from the video sequence using background subtraction or people detectors. Due to many reasons (background clutter, occlusions, etc.), the person silhouette can be badly aligned at different location of the extracted sub-images, or may even be truncated (missing occluded part, bad background subtraction segmentation, etc.). These issues may compromise the signature extraction as long as the learning of variations of a given body part feature requires to localize it at the same position on input images. Some approaches try to deal with this issue using a pairing step between images of a same person in term

silhouette localisation. They perform this pairing using some features matching increasing computation time.

We propose a fast method to perform the pairing step before signature extraction. This method is based on a fast color region comparison in Lab space, and it is detailed in the chapter 6.

- Despite they claim their robustness against people orientation variation, it seems that all the cited approaches were evaluated on datasets containing manually annotated images showing the same side of each person or with a small range of rotations in each camera. In the case of a person with a full rotation in the scene, all his/her side images are acquired. If the appearance of front side differs greatly from his/her back side (due to a backpack for example), a single signature for his/her appearance may be insufficient to allow his/her re-identification. Two cases may occur with the presented approaches: either only few features are kept if the approach is based on a selection of stable features, providing poor signature, or the variation of each feature are encoded in a large model, which may increase the re-identification failure probability due to less discriminative signature.

  We propose a context-aware approach, using spatio-temporal information, to divide a person signature into several sub-signatures corresponding to different visible sides. The processing time is not increased significantly as long as the signature extraction is performed using the same approach as if no signature subdivision is done. The sub-signatures are labelled according to the viewed side from which they are computed, allowing a faster matching process, while taking into account possible error of labels.

3

Overview of the Proposed Approach

In this manuscript, we propose a complete framework for people re-identification. The proposed approach includes three successive tasks: people detection, people tracking in mono-camera context and finally, visual signature extraction and comparison for re-identification. This chapter presents a general description of these three steps. The details of these tasks can be found in the next three chapters.

3.1 People Detection

The objective of this detection is to provide reliable inputs for both mono-camera people tracking and for people re-identification tasks (see figure 1.8). Detected persons on video sequences will be used as target initialization for mono-camera tracking in some specific cases, and as confirmation for other cases which will be more explained in mono-camera object tracking chapter (See chapter 5). They will also be used as request or candidates definition for re-identification task in both cases: static images and video sequences. The aim of separating these two cases is related to the type of used re-identification approach. According to some constraints which will be discussed in the chapter 6 concerning people re-identification, this task can use single-shot or multiple-shots to compute the visual signature of any person of interest or any candidate.

Like the most of state of the art approaches for people detection, the proposed approach follows the same global schema which consists in two main and separate steps: classifier training and people detection.

The proposed approach strongly optimize a state of the art approach proposed by Tuzel et al. [Tuzel 2007] and improved by Yao et al. [Yao 2008]. It keeps the original and improvements steps, but introduce an additional processing step in the training stage which speeds-up strongly both training and detection phases and improves detection performances.

A short discussion about the generalizability of the proposed optimization and its possible application on any training approach based on boosting to learn a cascade of classifier, regardless of the type of used features, is provided at the end of chapter 4

An overview of the proposed people detection algorithm is presented in the following sections. It contains the three main parts of people detection approaches: the pertinent features extraction, the classifier training and finally the candidate region selection for detection.

3.1.1 Pertinent Feature Selection: Region Covariance Descriptor

To highlight the improvements provided by our optimization method to people detector approaches, we have taken Tuzel et al. [Tuzel 2007] approach as baseline, and we have shown the improvements of our proposed approach by comparing our results with those of [Tuzel 2007] and also with those of [Yao 2008] which proposed other improvements to the original method, and with several other state of the art detectors.

The used features in these two approaches [Tuzel 2007, Yao 2008] and thus in our approach is Region Covariance Descriptor. Region covariance descriptor is a powerful way to encode a large amount of information inside in a given image region. It allows the encapsulation of a large range of different features in a single structure, representing the variances of each feature and the correlation between features.

Our choice to base our work on the approach of [Tuzel 2007] is partly due to this descriptor. In fact, Region Covariance Descriptor differs from the other local descriptors in two main ways.

First, Region Covariance Descriptor can be considered more as a generic container for various features, with a powerful bag of mathematical tools than a “rigid” local descriptor like SIFT [Lowe 2004], SURF [Bay 2008], HOG [Dalal 2005], LBP [Ojala 1996] and other local descriptors. In fact, while each local descriptor uses specific image information in a specific way, imposing a kind of rigidity to it, region covariance descriptor allows to use a large set of “basic” features (large amount, as there is more extractable basic features for each pixel). For example, both SIFT and HOG are based on image gradients only while LBP uses the values of direct neighbouring pixels of each pixel of interest. Some authors have combined some features to create hybrid features, by concatenating different descriptors for example but this kind of technique does not provide

the relationship between the different features belonging to the hybrid one. In contrast, Region Covariance Descriptors allow to uses several features in the same structure, and provide the information about the relationship between these features.

The second main difference is the scope of the descriptors or in other terms, the area of image that they describe. Local features generally describe areas with constant sizes. For example, unitary SIFT descriptor, in its standard version, is computed on a square region on 16x16 pixels and provide, after specific subdivisions on sub-regions of 4x4 pixels and bins separations on 8 bins, a feature vector of 128 dimensions. Of course, it is possible to enlarge or reduce the computing window size, but it requires important modifications, either in the subdivision method or in the resulting descriptor size. It also requires to ensure some constraints concerning the size of computing window and its square shape: for example, it is impossible to divide a square region of 17x17 equivalently or to use a rectangular window without losing some robustness to rotation in comparison to the standard version. In contrast, a region covariance descriptor can describe a very small region of few pixels in the same way that it can describe the whole image, without any necessary adaptation, and independently of the shape of the region (square or rectangular one), providing descriptors with same structures and dimensions. In fact, the dimension of a covariance descriptor relies to the size of the used feature set and not to the size of the described region.

Due to these advantages, region covariance descriptors will be used in the re-identification part by changing only the used basic features.

More details and explanations about Region Covariance Descriptors and their advantages are provided in chapter 4.

3.1.2 Classifier Training: Cascade of Classifiers Using LogitBoost Algorithm in Riemannian Manifold

Our main contribution takes place in this processing step. The proposed approach uses an adapted form of LogitBoost algorithm to train a cascade of classifiers for people detection. Due to the fact that region covariance descriptors relies to Riemannian Manifolds, the original LogitBoost algorithm was modified by Tuzel et al. [Tuzel 2007] to deal with this specificity.

Like all state of the art approaches which are based on an off-line training of a binary classifier to separate persons from non-person images, two type of images datasets are used: a set of positive images corresponding to various persons with different appearances and clothes to ensure the genericity of the classifier and a set of negative images corresponding to various things except people. This second dataset must be as various as possible and in term of content, and is supposed to be many times larger than the

positive dataset.

In the chapter 4, the training algorithm will be entirely detailed. The trained cascade of classifiers consist in a set of ordered strong classifiers, each of which consists in a set of weak classifiers. We will see that to train one weak classifier, heavy computations are required. The processing time required for training one weak classifiers is directly proportional to the number of randomly selected weak learners, but also and in large extent to the number of used samples. It is then clear that the larger the training dataset is, the slower the training is. At the same time, the more various and numerous the negative training data are, the less false positive detection rate is.

We propose a pre-training step which allows the use of large training negative datasets while it decrease significantly training time in a first stage and it provide a better cascade of classifiers in term of structure and content in a second stage, speeding-up the detection also. This step consists in clustering negative data before training in a specific way, to train each cascade level with a given cluster, specializing the corresponding strong classifier to reject a specific kind of image information.

The other main reason of our choice to take Tuzel et al. [Tuzel 2007] as a basis for our work lies in the nature of the trained classifier. Training a cascade of classifiers using a LogitBoost algorithm in Riemannian Manifolds is an interesting way to show that our method can be generalized to any kind of object detector which is based on cascade of classifier, due to the fact that both vector space and Riemannian manifolds are used during the training and in for method.

3.1.3 Candidate Regions Selection for People Detection: Dense Searching, Real World Candidate selection and Background Subtraction Filtering

Three methods for candidate regions selection are used. Dense searching method is the standard and the usable method in any case because it does not requires any additional external information. It can be applied on any image, but it is also the slowest method due to the highest number of tested candidate regions.

Whether camera calibration information is available, the candidate region selection is focused on the areas where it is more likely to find people, i.e. all areas of the image touching the ground floor in the real world. If scene context information is available too, this targeting is more precise by avoiding all areas that can not contain people due to the presence of scene static objects (Walls, buildings, etc.).

The last selection method, which is mainly used for our mono-camera tracking algorithm, is based on motion regions targeting. This method can be used in both previously

mentioned methods, as long as it is applied on video sequences. It is more a filtering method than an independent one for candidate regions selection. It consists in focusing the detection only on moving objects, detected by background subtraction algorithm, and using camera calibration to reduce scale and width/height ratios of searching windows. Using this filtering method reduces strongly detection time, but can avoid some detection in case of static people on images.

Due to the context of our study and the work hypothesis concerning the use of static and calibrated cameras, all the possible candidate regions are pre-computed once before the detection or during the detection on the first frame of the input video sequence. These candidate regions are stored and used on each new frame where motion is detected, avoiding the necessity to calculate the 3D-2D projections at each frame.

3.2 Mono-Camera Object Tracking

In our study, the objective of tracking people in mono-camera context consists in two main points: it allows to extract and to learn robust appearance-based visual signatures and it provides useful information for spatio-temporal filtering step during people re-identification.

To re-identify people in a camera network using appearance matching, it can be sufficient to detect all appearing persons on all the frames of available video sequences using a people detector, and extract single-shot visual signatures for each detected person. However, this method increases strongly the number of possible candidate as it considers each detection as a different person, due to the non-use of temporal information. Generally, it is better to use multiple-shot based visual signature for many reasons which are detailed in the re-identification chapter (see chapter 6). Using multiple images of a person to compute its visual signature requires to detect him on several frames on the video sequences provided by one camera, and to ensure that all these detections represent the same person. Mono-camera person tracking is the best way to perform this task.

We also use person tracking information to robustify our visual signatures by partitioning it in sub-visual signatures according to the visible side of a person in a given camera. In fact, the motion direction of a person with respect to a given camera can be easily extracted and used to infer if that this person is seen from front/behind or in profile, so the extracted visual information can be assigned to the right person position and allow more precise comparison during re-identification process. All these aspects are more discussed in the chapter 6.

Finally, mono-camera people tracking with calibrated cameras provides also useful

information for spatio-temporal coherency filtering in both cases of overlapping and non-overlapping fields of view between the cameras. In case of overlapped fields of view, only people in the neighbourhood of a given person at a given time are considered as candidates for re-identification. In case of non-overlapped fields of view between cameras, using buildings plans or city maps or any contextual information, in addition to camera calibration information, allows to avoid incoherent and unlikely candidates in some cameras for a targeted person in an other camera, by checking distance/velocity coherency or possible trajectories.

We propose a mono-camera "object" tracking framework based on SIFT features (Scale-Invariant Feature Transform) [Lowe 2004] tracking by particle filtering. It mainly use detected moving object by background subtraction as target initialisation and candidates providing, and in some cases which will be discussed in chapter 5, the people detector we present in chapter 4. A data association method, based on the reliability of tracked SIFT features, is proposed to pass reliably from point tracking to object tracking to create the temporal links between objects of each frame and thereby provide objects trajectories. This data association method also allows to detect partial and full occlusion situations, which are managed by our proposed "fast occlusion managment approach". This last task is performed using several information: SIFT Features matching, dominant colors descriptor and "real world" object information like size and velocity (provided thanks to camera calibration).

This framework is a generic object tracking algorithm which can track any kind of objects. People tracking is performed by targeting only people among detected moving persons, either by using the "real world" information of the object (especially the real dimensions), by applying the people detector or by using both. All the cases are discussed in the chapter 5. It is also generic in term of used features for tracking as long as the object model consists in sparse local descriptors. In fact, in our presented work, we use SIFT features for this task, but any interesting local descriptor can be used as long as it ensures invariance and robustness to classical challenges.

3.2.1 Object Detection: Background Subtraction VS. People Detection

Our object tracking algorithm performs by creating temporal links between detected/tracked objects on the frame at time "t-1" with their corresponding objects detected on the frame at time "t". It means that objects have to be detected at each frame in a fist stage. At time "t", some new detected objects can appear for the first time, other objects can disappear by leaving the field of view or by being occluded. All possible cases are managed in the data association step.

This object detection is performed by a state of the art background subtraction al-

gorithm. It provides all moving objects delimitations on the images. These moving objects can be various things: objects of interest (people, cars, etc.), other objects (tree branches, flags, etc.), illumination change and noise.

Focusing on a given type of objects can be performed using camera calibration information, by checking the estimated real world dimension of any detected object and its velocity for example and verifying that it correspond to known value ranges. This method provides good results in simple situations, but it fails when some complicate cases occurs: shadows, grouped people, closest objects, etc.

We propose to use the presented person detector in collaboration with background subtraction to focus on people by applying it on the moving image regions.

3.2.2 SIFT Features Detection and Selection

In the proposed approach, tracked object are modelled by a set of point features. This choice is justified by two main reasons: the flexibility of points representation of an object allow to deal with objects deformations and rotations easily since points are independent. The object occlusions can be detected easily by noting the disappearance of a part or all used points.

Once one target object detected, we model this object by a set of SIFT features. A SIFT feature [Lowe 2004] consist in a SIFT point of interest, detected following some conditions, around which a SIFT descriptor is computed in specific way. This descriptor is assigned to the point of interest.

The several steps of SIFT feature computing, which will be recalled in chapter 5, ensure their robustness and invariance. The optimal algorithm parameters and thresholds provided in [Lowe 2004] ensure the best robustness and invariance level. However, these parameters are too restrictive to ensure a minimal representation of object of interest with enough amount of points, and with a whole covering. Small and noisy object images provided by many video surveillance systems may be fully or partially empty in term of SIFT points.

The proposed approach changes some SIFT points detection parameters (curvature and contract thresholds) and provide more detected points. This allows to ensure a sufficient number of points on all the object image, even if the robustness of these points is decreased. The reliability decreasing of SIFT points is compensated by the proposed general data association frameworks witch use reliability measure on tracked SIFT features to create the temporal links between objects.

The approach filters the detected SIFT points using a grid subdivision of the object image. A uniform distribution of SIFT features on the whole image is then provided by this method.

3.2.3 SIFT Features Tracking by Particle Filtering Method

Due to the cited advantages of particle filtering methods for tracking (see sec. 2.2.2.2), the proposed approach track detected and selected SIFT points using a customized particle filter. Each SIFT point is tracked using a set of particles. The point tracking is performed by the two standard steps which are perdition and update.

Each SIFT point movement is characterised by an adaptive dynamic model, based on SIFT points position and velocity. Our approach use a linear regression function, computed on a definite number of the last positions of the SIFT point, to adapt the motion model. By this way, movement velocity and direction changes are well handled.

The prediction step for each SIFT point tracking is performed by applying its motion model to all associated particles, including a random noise estimation to each particle location projection.

Once all particles are projected in the current frame, the update (or correction) step provide the estimation of the current location of the tracked SIFT point as the centroid of all its particles. Before new SIFT point location estimation, its particles are weighted, sampled and resampled to make them focus on the most likely positions for SIFT point. We use the Importance Sampling Resampling method due to its ability to avoid information degeneration (see section 2.2.2.2). Our method propose a new hybrid weighting method for particles.

3.2.3.1 Hybrid Particles Weighting for Sampling Resampling Step

Most of state of the art particle weighting in object tracking particle filters methods are based on a similarity measure between the used descriptors. The proposed method keeps this weighting technique since the similarity measure is an important information, but due to the addressed context (video surveillance), the low resolution of image, the small size of objects of interest, and the low contrast with background may greatly alter this similarity measure. A practical example is provided in chapter 5. To avoid this risk, a second information type is used for weighting.

The proposed approach combine background subtraction result with the similarity measure in an effective way to ensure a better weighting process. The background subtraction result is used in a continuous way even if it is provided in binary form. Our hybrid weighting method has the other advantage to be robust to background subtraction algorithm performances. It is based on the moving pixel density calculation, and is explained in chapter 5.

3.2.4 Data association: Temporal Links Creation

The previous steps allow to track SIFT points independently. To infer the object of interest movement and localization, a data association step, based on tracked SIFT points, is performed.

Temporal links between objects of successive frames are built depending on the localisation of tracked SIFT points on the currently detected object, weighted by their reliability measures. This method is faster than a basic detected points matching.

The proposed approach allows to detect all tracking situations: single objects moving, grouping object, separation of group of objects, and occlusions. This is performed using the tracked SIFT points localization and visibility on current frame. This method is detailed in the chapter 5.

We propose a method to build these temporal links in a reliable way, using all available information from SIFT features: (1) The spatial repartition of SIFT points to detect occlusions and to delimit the object and (2) the reliability of SIFT features to weight all the possible links between tracked objects at time "t-1" and the detected objects at time "t"

3.2.5 Fast Occlusion Management

The proposed data association method allows to detect occlusion situations thanks to point representation and tracking.

Partially occlusions are handled by a continuous tracking process on the remaining visible SIFT points. If the partially occluded object became fully visible, new additional SIFT points are detected and assigned to the tracked object using the same process as the one described for SIFT points detection and selection. This is due to maintain a global and well distributed representation of the object

Full occluded objects are handled differently. If a tracked object is lost due to a full collusion, it is stored as an "object to reaquire".

During the tracking process of a fully visible object, some additional features are extracted and used to "learn" a coarse model for reacquisition purpose only (not for tracking). This model consists in a combination of the variation of real world object dimension and velocity extracted thanks to camera calibration in addition to the extracted dominant color descriptor (the two main colors are considered in our approach).

A matching between the last visible set of SIFT point of occluded object (taken before occlusion starting) and those of a non linked candidate object in current frame is also performed. We can afford this computational time consuming task since it is not performed in each tracking iteration (on each video frame), but on specific cases.

Note that this last matching method is used as a validation-only criteria. If a candidate object present a high matching rate of matching SIFT points with the previously occluded one, and according to defined thresholds, this candidate object can be considered as the required one of the previously occluded object. Otherwise, if SIFT point matching fails, the candidate point is not rejected. The previous cited model for reacquisition is used to test object matching. This is due to the fact that occluded object may reappear in another pose or orientation with respect to camera, hiding its previously used SIFT points. The object model for reacquisition is a more general one witch can be used even if object pose or orientation change during the occlusion.

Once a previous fully occluded object have reappeared and reacquired, new SIFT points are detected and assigned to the tracked object using the same process as the one described for SIFT points detection and selection.

Finally, non linked detected objects in the current frames, either by continuous tracking or by reacquisition, are used to initialize new object tracking since they are considered as new appearing objects in the scene.

3.3 People Re-identification

People re-identification is the final aim of the proposed framework. The proposed approach is based on [Farenzena 2010] approach which provides good performances and which performs in real-time (or in pseudo-real time if the number of considered images per person is larger). Some of the baseline issues and more general state of the art approach issues are identified and some solutions are proposed to solve them, improving the performances of the approach, making it more generic (no off-line training or video-operator interaction are required), while maintaining low processing time requirements.

The initial approach of [Farenzena 2010] performs in three steps:

First, human body image is divided into four parts using two asymmetry axes which separates vertically the head from the torso and the torso from the legs, and two symmetry axes which separate horizontally torso into two parts, and legs into two parts also. The head part is ignored in this approach because of low amount of information it provides in general case (large scale video surveillance). This body subdivision is performed using both foreground pixels separation and their color values.

Once the human body separation performed, three different features are computed on each body part: Weighted color histograms (WH) in HSV space, Maximally Stable Color Regions (MSCR) and Recurrent High-Structured Patches (RHSP). The final signature consists of the combination of these features.

Finally, the comparison between two signatures is performed by computing a dissimilarity measure between them. This dissimilarity measure is the sum of weighted dissimilarity measures between each pair of similar features (WH). The weights are fixed once by an off-line learning step on a subset of images from VIPeR dataset.

In the following paragraphs, the main identified issues of this baseline approach (and some other more general issues) as the proposed solutions are briefly exposed.

3.3.1 Dependency of Visual Signatures to People Orientations

This is a general issue for many state of the art appearance-based approaches for people re-identification. People appearance may probably be different on its several sides. Opened jackets with different color than the t-shirt may provide different colors whether it the person is observed from his/her frontal or back side. Backpacks introduce erroneous information on the back side of people when they are not visible when the people are observed from front side. Some textures and patterns which are on front side of cloths are not observed if the person in observed from back or in profile. Many other examples can illustrate the importance of people visible side identification for visual signatures computation and comparison.

We propose method to classify people images according the their visible side, using real world information, provided by our mono-camera tracking algorithm. The people trajectories are segmented according their walking directions. Assuming that people walk forward, their trajectories indicate their orientation and thereby, their visible side with respect to the cameras, thanks to some processing steps using camera calibration information. These processing steps are detailed in chapter 6. The unique visual signature is replaced then by a set of sub-visual signatures assigned to each visible side class. This classification method is accompanied by the corresponding signature comparison method, dealing with the all possible cases (common visible side classes, adjacent visible side classes, neither common nor adjacent visible side classes).

3.3.2 Unreliable Body Subdivision + Images Alignment Issue for Multiple-shot Case

The proposed method for human body subdivision (symmetry and asymmetry axes estimation) is mainly dependent on the quality of background subtraction and to the contrast between people and the foreground. Generally, the estimated axes are not correct but more or less inaccurate. For small shifting of the asymmetry axes with respect to their real localization, the final visual signature alteration is negligible. For this reason, we have decided to avoid this processing step and to use statistical subdivision of human

body, by positioning the head-torso separation line at 1/5 of the human height (starting from the top) and the torso-legs at 3/5 of the human height.

This solution may cause an important issue: it is dependent on the quality of people delimitation. People who are not well centred on the used images (or bounding boxes) or who have some missing parts (cropped head, feet, etc.) due to bad background subtraction or people detection results provides bad statistical body subdivision. For this reason, and to improve part-to-part comparison, we propose a fast method to align correctly all the images of a given person. This allows to correct some images by shifting people to the center and by removing background in images margins, and also to remove images with significant error (cropped important parts). This image alignment is performed using fast browsing of superposed images using Lab color space for similarity maximization. Some processing steps like image downscaling and camera calibration information use speed up the alignment processing.

3.3.3 Exclusive Use of Unnormalized Color Information

The initial approach of [Farenzena 2010] uses images directly without managing the difference in color rendering between cameras issue. We propose to use color histogram equalization to minimize the effect of this issue. We prefer this kind of approaches instead of colorimetric camera calibration due to the issues and constraints of this last approaches, which consist of non-bijectivity of transfer functions and the complexity of their application in large scale video-surveillance system.

On the other hand, [Farenzena 2010] characterize a visual signature exclusively using color information. Weighted color histograms and Maximally Stable Color Regions are fully color-based features. Recurrent High-Structured Patches, even they are selected using texture information (patches entropy and LNCC maps), their final characterization is performed using simple color histograms too.

We propose to replace the simple color histograms which characterize RHSP features by covariance descriptors which are built using both color and texture information, improving the discriminative power of RHSP features.

To use additional texture information, we propose to add tracked SIFT features the the final signature too. All the required modifications in the initial approach to manage this add are performed (modification of the signature comparison by adding a weighted dissimilarity measure of SIFT features in the final dissimilarity measure between signatures).

Both SIFT and RHSP features being local descriptors, their use is controlled and managed by the adaptive feature weighing system we propose and which is described in the next paragraph.

3.3.4 Fixed Weights for Each Descriptor

In the baseline approach of [Farenzena 2010], The weight of each feature in the final signature comparison is fixed after an off-line learning on a part of the used dataset. This poses the three following issues:

First, the off-line learning of the best weights is not conceivable for our work since we try to provide a generic and turnkey system for people re-identification, which does not require any external interaction after deployment.

Second, the used weights are the same for the whole dataset. We believe that this kind of weighting method is not the best one. In fact, for each person to re-identify, the system may focus more on the most discriminant information in this person appearance. A given person appearance may be highly rich in terms of color while it is poor in terms of textures (uniform large color regions on the cloths for example) and another person appearance in the same dataset (or in the same camera network) may be poor in terms of color (a unique dark color for example) while it is rich in terms of textures (created by another unique color). These two persons may not be compared with candidate people using the same features weights. The first person re-identification may focus more on color information while the second person re-identification may focus more on texture, event whether both people are in the same dataset or observed by the same camera.

Finally, using the visible side classification, the local nature of SIFT and RHSP features implies the integration the people orientations in these two feature weighting. It means that SIFT and RHSP feature weights vary greatly for the same query person according the compared visible sides.

To deal with these issues, we propose an adaptive weighting method, to assign the adequate weight to each feature of the visual signature, according to the richness/discriminative power of each type of information (Color and Texture) and to the considered visible sides. The assigned weights are then not dataset/camera network related, but visible side and information richness/discriminative power based for each person. The heterogeneity of the computed distances for the several people re-identifications is not an issue since the aim is to find for each query person, the most likely corresponding candidate, independently on the other re-identification queries.

4

EFFICIENT PEOPLE DETECTOR BASED ON COVARIANCE DESCRIPTORS

This chapter describes in detail the proposed people detector, mainly based on Tuzel et al. [Tuzel 2007] approach and its improvements by Yao et al [Yao 2008]. This people detector is based on a cascade of classifiers, trained using LogitBoost algorithm on Region Covariance Decriptors to detect full human body on both static images and video sequences. The process is performed in two separate steps: an offline training step, and a classification step.

4.1 Region Covariance Descriptor

Region covariance descriptors are a powerful way to encode a large amount of information inside a given image region. Unlike the concatenation of several feature vectors, which provide a final vector containing independent feature information, a Region Covariance Descriptor allows the encapsulation of a large range of different features in a single structure, representing the variances of each feature in the represented image region and the correlation between these features.

Tuzel et al. [Tuzel 2006] first introduce the use of covariance matrices as a descriptor for object classification.

Let I be an image of dimension $W \times H$. We can extract at each pixel location $\mathbf{x} = (x, y)^T$ a set of d features such as intensity, color, gradients, filter responses, etc.

For a given rectangular region R of I, let $\{z_k\}_{k=1..S}$ be the d-dimensional feature points inside R. The region R is represented with the $d \times d$ covariance matrix of the feature

points

$$C_R = \frac{1}{S-1}\sum_{k=1}^{S}(z_k - \mu)(z_k - \mu)^T \tag{4.1}$$

where μ is the mean vector of the points z_k and S the number of pixels within R.

The diagonal entries of covariance matrix represent the variances of each feature, and the non-diagonal entries are their respective correlations.

4.1.1 Fast Covariance Computation Using Integral Images

A large number of covariance descriptors are required to achieve the training of classifier cascade and for an effective process. The computation of all the feature sums, means and variances for each region has a high cost in term of processing time. To deal with this issue, Integral Images are ideally suited to minimize the number of numerical operations.

Integral Images are intermediate image representations used for the fast calculation of region sums [Simard 1999, Viola 2001]. Each pixel of the integral image is the sum of all the pixels inside the rectangle bounded by the upper left corner of the image and the pixel of interest. For an image (i), the Integral Image value at a pixel coordinates (x, y) is given by:

$$II(x, y) = \sum_{\substack{x' \leq x \\ y' \leq y}} i(x', y') \tag{4.2}$$

The Integral Image of any image can be computed efficiently in a single pass over the considered image, using the fact that the value in the summed area table at (x, y) is just:

$$II(x, y) = i(x, y) + II(x-1, y) + II(x, y-1) - II(x-1, y-1) \tag{4.3}$$

The main interest of Integral Image representation is the possibility to compute the sum of pixels in any rectangular region of the image (see figure 4.1) using a constant number of 4 access and 3 simple mathematical operations:

$$\sum_{\substack{x_1 < x' \leq x_2 \\ y_1 < y' \leq y_2}} i(x', y') = II(x_2, y_2) + II(x_1, y_1) - II(x_1, y_2) - II(x_2, y_1) \tag{4.4}$$

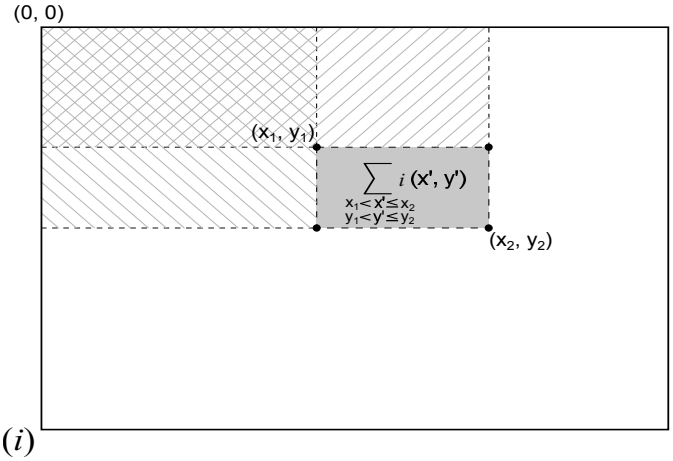


Figure 4.1: Fast computation of the sum of pixels inside any rectangular region, using Integral Image representation.

Integral Images can be used to reduce greatly Region Covariance Descriptors computation. From equation 4.1, any entry $C_R(i,j)$ of the covariance matrix can be written as:

$$C_R(i,j) = \frac{1}{S-1}\sum_{k=1}^{S}(z_k(i) - \mu(i))((z_k(j) - \mu(j)) \tag{4.5}$$

This equation can be rewritten as:

$$C_R(i,j) = \frac{1}{S-1}\left[\sum_{k=1}^{S} z_k(i)z_k(j) - \frac{1}{S}\sum_{k=1}^{S} z_k(i)\sum_{k=1}^{S} z_k(j)\right] \tag{4.6}$$

The covariance matrix in a given rectangular region R can be computed by firstly computing the sum of each feature dimension $z(i)_{i=1\ldots d}$ as well as the sum of the multiplications of any two feature dimensions $z(i)(j)_{i,j=1\ldots d}$.

An integral image $P(x,y,i)$ is computed for each sum of each feature dimension i and an integral image $Q(x,y,i,j)$ is computed for each sum of the multiplication of any two feature dimensions i and j

Due to the symmetric nature of covariance matrices, only upper (or lower) triangle values are computed, it means that d Integral Images $P(i)_{i=1\ldots d}$ and $(d^2+d)/2$ Integral Images $Q(i,j)_{i,j=1\ldots d}$ are computed

Considering any rectangular region $R(x_1,y_1,x_2,y_2)$ of image, and using equation (4.6) and equation (4.4), it is possible to compute the associated covariance matrix in a constant number of operations as:

$$C_{R(x_1,y_1,x_2,y_2)} = \frac{1}{S-1}\Big[Q_{x_2,y_2} + Q_{x_1-1,y_1-1} - Q_{x_2,y_1-1} - Q_{x_1-1,y_2} - \frac{1}{S}(P_{x_2,y_2} + P_{x_1-1,y_1-1} - P_{x_2,y_1-1} - P_{x_1-1,y_2})(P_{x_2,y_2} + P_{x_1-1,y_1-1} - P_{x_2,y_1-1} - P_{x_1-1,y_2})^T\Big] \quad (4.7)$$

4.1.2 Used Features

For people detection purpose, we use an initial set of 8-dimensional set for features, like in [Tuzel 2007] and in [Yao 2008]. Tuzel et al. [Tuzel 2007] use the following set of pixel features (see figure 4.2):

$$\left[x \;\; y \;\; |I_x| \;\; |I_y| \;\; \sqrt{I_x^2 + I_y^2} \;\; |I_{xx}| \;\; |I_{yy}| \;\; \arctan\frac{|I_x|}{|I_y|}\right]^T \quad (4.8)$$

where:

x and y are the pixel coordinates

I_x and I_{xx} are respectively the first and the second order intensity derivatives on X axis

I_y and I_{yy} are respectively the first and the second order intensity derivatives on Y axis

$\sqrt{I_x^2 + I_y^2}$ and $\arctan\frac{|I_x|}{|I_y|}$ are respectively the magnitude and the orientation of the gradient (the edge).

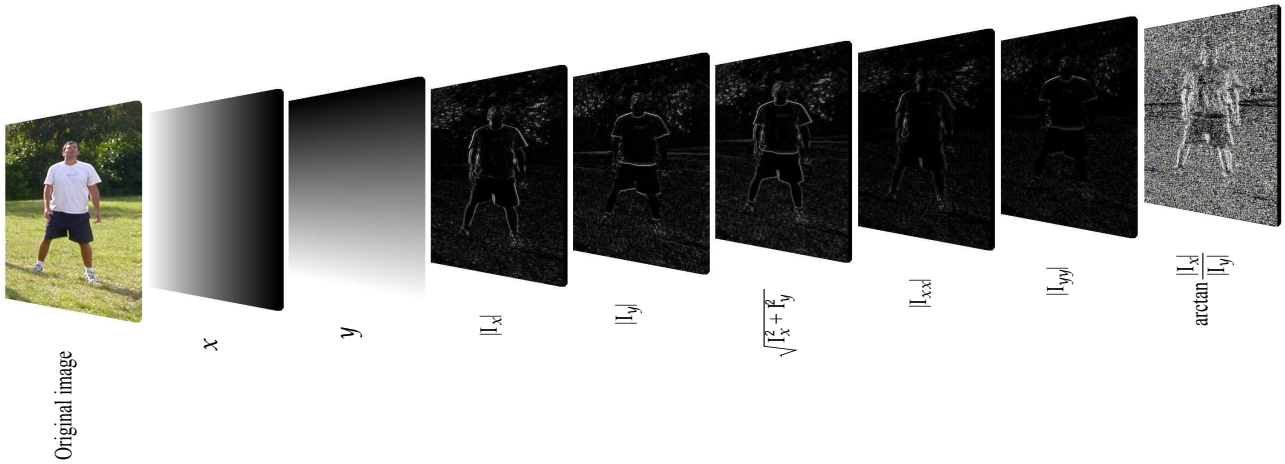


Figure 4.2: The 8 pixel features used in [Tuzel 2007].

Yao and Odobez [Yao 2008] replace the two second derivatives features $|\mathbf{I}_{xx}|$ and $|\mathbf{I}_{yy}|$ by two foreground measures $\mathbf{G}$ and $\sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}$. $\mathbf{G}$ denotes the foreground probability value (a real number between 0 and 1 indicating the probability that the pixel x belongs to the foreground), and $\mathbf{G}_x$ and $\mathbf{G}_y$ are the corresponding first order derivatives. These foreground features are obtained using a background subtraction technique which is restricted to moving people (see figure 4.3). In the context of human detection in videos, foreground measures should be much more informative.

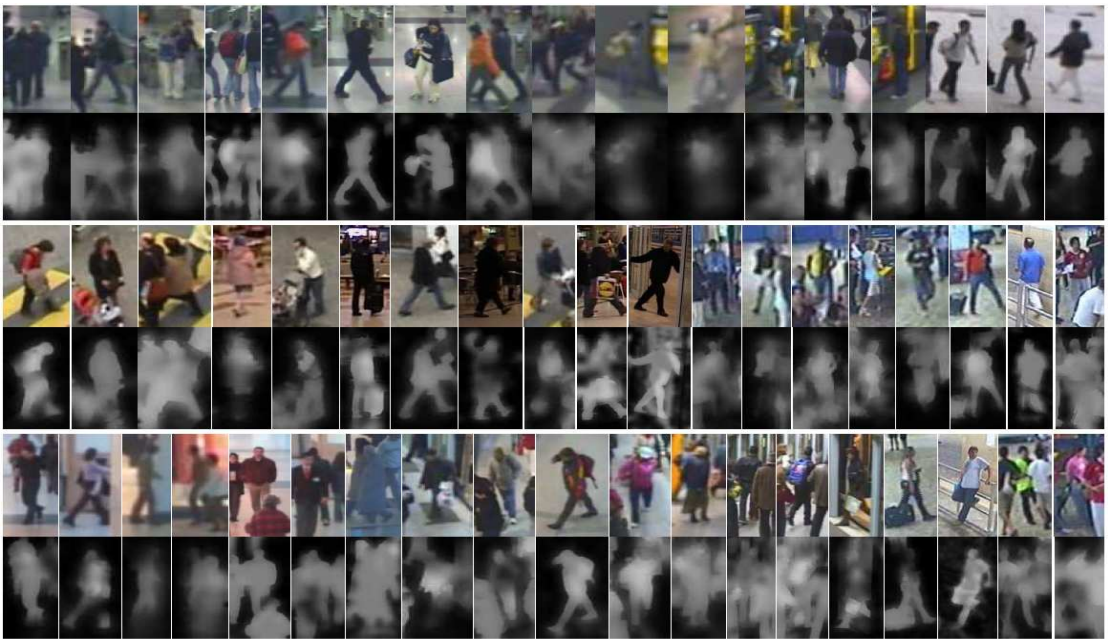


Figure 4.3: [Yao 2008] positive examples with corresponding foreground probability maps (light - high probability, dark - low probability).

Despite the improvements provided by the use of these two background features instead of the two second order derivative ones, we have used both set of features. The final evaluation and comparison of our people detector is performed using [Yao 2008] features but some necessary tests and comparisons to highlight our intermediate contribution are performed using [Tuzel 2007] features set. This is due to the availability of more details concerning the trained classifier in [Tuzel 2007]: the cascade of classifiers structure, the required training time, the negative rejection rate per cascade level, the average detection time per image, etc. so to make sense to the comparison, it has to be performed in the same conditions.

In [Yao 2008], only global evaluation results and their comparison with state of the art ones are provided.

Note that in the considered case of 8-feature set, the covariance matrix will contain

36 different values (due to the symmetry), and 44 Integral Images are computed to speed up the computing process (8 integral images for the representation of the sums of each feature independently and 36 for the representation of the sums of product for each pair of features).

4.1.3 Covariance Normalisation

The covariance features are robust towards constant illumination changes. To enhance the robustness against local linear variations of the illumination on a given subregion r with respect to its largest containing region R, a normalization step is performed on the covariance matrix.

First, both covariance matrices $\mathbf{C_r}$ and $\mathbf{C_R}$ are computed using integral representation. The values of covariance matrix $\mathbf{C_r}$ are normalized with respect to the standard deviations of their corresponding features inside the containing region R as:

$$\hat{C}_r = diag(C_R)^{-\frac{1}{2}} \cdot C_r \cdot diag(C_R)^{-\frac{1}{2}} \tag{4.9}$$

where $\hat{C}_r$ is the normalized covariance matrix of the subregion r, C_r its initial covariance matrix, C_R the covariance matrix of the containing region and $diag(C_R)$ is a matrix equal to C_R at the diagonal entries and zero value at all other entries.

4.2 Region Covariance Descriptors as Riemannian Manifold

A manifold is a topological space which is locally similar to an Euclidean space. Every point on the manifold has a neighborhood for which there exists a homeomorphism (one-to-one, onto, and continuous mapping in both directions) mapping the neighborhood to $\mathbb{R}^m$. For differentiable manifolds, it is possible to define the derivatives of the curves on the manifold. The derivatives at a point $\mathbf{X}$ on the manifold lie in a vector space T_X, which is the tangent space at that point.

A Riemannian manifold $\mathcal{M}$ is a differentiable manifold in which each tangent space has an inner product $<,>_{X \in \mathcal{M}}$, which varies smoothly from point to point. The inner product induces a norm for the tangent vectors in the tangent space such as that $\|\mathbf{v}\|^2_{\mathbf{X}} = < \mathbf{v}, \mathbf{v} >_{\mathbf{X}}$.

The minimum length curve connecting two points $\mathbf{X_i}$ and $\mathbf{X_j}$ on the manifold is called the geodesic and the distance between the points $d(\mathbf{X_i}, \mathbf{X_j})$ is given by the length of this curve (see figure 4.4).

Let $v \in T_{X_i}$ and $X_i \in \mathcal{M}$. From X_i, there exists a unique geodesic $\gamma_v(t)$ starting with the tangent vector y. The exponential map $exp_{X_i} : T_{X_i} \longmapsto \mathcal{M}$ maps the vector

y to the point reached by this geodesic, and the distance of the geodesic is given by $d(X_i, exp_{X_i}(v)) = \|\mathbf{v}\|_{\mathbf{X_i}}$. The inverse mapping is defined by $log_{X_i} : \mathcal{M} \longmapsto T_{X_i}$.

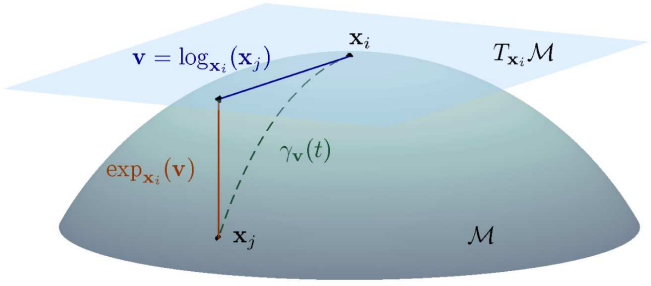


Figure 4.4: A two-dimensional manifold. The tangent plane at X_i, together with the *exp* and *log* maps relating X_i and X_j are shown. (source [Goh 2008]).

In the following paragraphs, points in the vector space are noted with small bold letters while points on the Riemannian manifold are noted with capital bold letters.

The covariance matrices are symmetric positive definite matrices. The set of $d \times d$ dimensional symmetric positive definite matrices, noted Sym_d^+ can be formulated as a connected Riemannian manifold, and an invariant Riemannian metric on the tangent space of Sym_d^+ is given by [Pennec 2006]

$$< \mathbf{y}, \mathbf{z} >_{\mathbf{X}} = trace(\mathbf{X}^{-\frac{1}{2}}\mathbf{y}\mathbf{X}^{-1}\mathbf{z}\mathbf{X}^{-\frac{1}{2}}) \tag{4.10}$$

The two mapping functions are:

$$exp_{\mathbf{X}}(\mathbf{y}) = \mathbf{X}^{\frac{1}{2}} \exp(\mathbf{X}^{-\frac{1}{2}}\mathbf{y}\mathbf{X}^{-\frac{1}{2}})\mathbf{X}^{\frac{1}{2}} \tag{4.11}$$

$$log_{\mathbf{X}}(\mathbf{y}) = \mathbf{X}^{\frac{1}{2}} \log(\mathbf{X}^{-\frac{1}{2}}\mathbf{y}\mathbf{X}^{-\frac{1}{2}})\mathbf{X}^{\frac{1}{2}} \tag{4.12}$$

The *exp* and *log* are the ordinary matrix exponential and logarithm operators. Not to be confused with $exp_{\mathbf{X}}$ and $log_{\mathbf{X}}$ which are manifold specific operators, and which are point dependent, $\mathbf{X} \in Sym_d^+$. The tangent space of Sym_d^+ is the space of $s \times d$ symmetric matrices, and both the manifold and the tangent spaces are $m = d(d+1)/2$ dimensional.

The ordinary matrix exponential and logarithm of a symmetric matrix can be computed easily using its eigenvalue decomposition. Let $\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$ the eigenvalue decomposition of a symmetric matrix. The ordinary *exp* and *log* matrix operators are given by:

$$\exp(\Sigma) = \sum_{k=0}^{\infty} \frac{\Sigma^k}{k!} = \mathbf{U} \exp(\mathbf{D})\mathbf{U}^T \tag{4.13}$$

$$\log(\Sigma) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} (\Sigma - \mathbf{I})^k = \mathbf{U} \log(\mathbf{D}) \mathbf{U}^T \tag{4.14}$$

$\exp(\mathbf{D})$ and $\log(\mathbf{D})$ are obtained by applying respectively exponential and logarithm functions on the diagonal entries of the diagonal matrix $\mathbf{D}$.

Note that the exponential operator is always defined whereas the logarithms only exists for symmetric matrices with positive eigenvalues, Sym_d^+, which is the case for the considered covariance matrices.

From the definition of the geodesic given above, the distance between two points on Sym_d^+ is measured by substituting (4.12) into (4.10)

$$\begin{aligned} d^2(\mathbf{X}, \mathbf{Y}) &= < \log_{\mathbf{X}}(\mathbf{Y}), \log_{\mathbf{X}}(\mathbf{Y}) >_{\mathbf{X}} \\ &= trace(\log^2(\mathbf{X}^{-\frac{1}{2}} \mathbf{Y} \mathbf{X}^{-\frac{1}{2}})) \end{aligned} \tag{4.15}$$

A minimal representation of the points (covariance matrices) in the tangent space are required for classification. Tuzel et al. [Tuzel 2007] define an orthonormal coordinate system for the tangent space with the vector operation. The orthonormal coordinates of a tangent vector $\mathbf{y}$ in the tangent space at point $\mathbf{X}$ is given by the vector operator:

$$vec_{\mathbf{X}}(\mathbf{y}) = vec_{\mathbf{I}}(\mathbf{X}^{-\frac{1}{2}} \mathbf{y} \mathbf{X}^{-\frac{1}{2}}) \tag{4.16}$$

where $\mathbf{I}$ is the identity matrix, and the vector operator at identify is defined as

$$vec_{\mathbf{I}}(\mathbf{y}) = \begin{bmatrix} y_{1,1} & \sqrt{2}y_{1,2} & \sqrt{2}y_{1,2} & \dots & y_{2,2} & \sqrt{2}y_{2,3} & \dots & y_{d,d} \end{bmatrix}^T \tag{4.17}$$

Given a set of points (covariance matrices), the tangent space used for the minimal representation defined above is tangent to the Riemannian manifold of covariance matrices at a specific point which is the mean of all the considered points (covariance matrices).

The mean of a set of point $\{\mathbf{X}_i\}_{i=1\dots N}$ on Riemannian manifold is defined as:

$$\mu = arg \min_{\mathbf{X} \in \mathcal{M}} \sum_{i=1}^{N} d^2(\mathbf{X}_i, \mathbf{X}) \tag{4.18}$$

where d^2 is the distance metric defined in (4.15).

The mean point can be computed iteratively using the following gradient descent procedure

$$\mu^{t+1} = \exp_{\mu^t} \left[\frac{1}{N} \sum_{i=1}^{N} \log_{\mu^t}(\mathbf{Xi}) \right] \tag{4.19}$$

A weighted mean can be computed similarly, by

$$\mu^{t+1} = \exp_{\mu^t}\left[\frac{1}{\sum_{i=1}^{N} w_i}\sum_{i=1}^{N} w_i \log_{\mu^t}(\mathbf{Xi})\right] \tag{4.20}$$

4.3 LogitBoost Algorithm on Riemannian Manifolds

The classification process is performed using a cascade of classifiers which is trained using a LogitBoost algorithm on Riemannian Manifolds like in [Tuzel 2007] to allow comparison of our cascade of classifiers with the one obtained in [Tuzel 2007]. As it was mentioned in section 2.1.3.3, the main difference between Adaboost an Logitboost resides in the way the weak classifier errors are computed and thereby the way the best weak classifier is selected at each iteration. AdaBoost minimizes an exponential loss function while LogitBoost minimizes a logistic loss.

4.3.1 Standard LogitBoost Algorithm on Vector Spaces

As presented in [Friedman 1998], let $\{(x_i, y_i)\}_{i=1...N}$ be the set of training samples, with $y_i \in \{0, 1\}$ and $x_i \in R^n$. The goal is to find a decision function F which divides the input space into the 2 classes.

In LogitBoost, this function is defined as a sum of weak classifiers, and the probability of a sample x being in class 1 (positive) is represented by

$$p(x) = \frac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}} \tag{4.21}$$

$$F(x) = \frac{1}{2}\sum_{l=1}^{N_L} f_l(x). \tag{4.22}$$

where $f_l(x)$ is the trained weak classifier at the l^{th} iteration.

The LogitBoost algorithm iteratively learns the set of weak classifiers $\{f_l\}_{l=1...N_L}$ by minimizing the negative binomial log-likelihood of the training data:

$$-\sum_{i}^{N}[y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))], \tag{4.23}$$

through Newton iterations. At each iteration l, this is achieved by solving a weighted least-square regression problem:

$$\sum_{i=1}^{N} w_i \|f_l(x_i) - z_i\|^2 \tag{4.24}$$

where
$z_i = \frac{y_i - p(x_i)}{p(x_i)(1-p(x_i))}$ denotes the response values,
and the sample weights are given by $w_i = p(x_i)(1 - p(x_i))$.

4.3.2 LogitBoost Algorithm on Riemannian Manifolds

The trained cascade consists of a list of ordered strong classifiers. Each strong classifier contains a set of weak classifiers. A weak classifier is defined by a subregion of interest inside the detection region, the corresponding mean value of covariance descriptors of all positive samples in the same subregion, and the corresponding regression function.

To train the LogitBoost cascade of classifier using covariance descriptors, the standard LogitBoost algorithm is not usable as it is. In fact, covariance descriptors do not belong to vector spaces but to the Riemannian manifold $\mathcal{M}$ of $d \times d$ symmetric positive definite matrices Sym_d^+.

Based on the previously presented invariant Riemannian metric, and the minimal representation of covariance matrices on the tangent space proposed in [Tuzel 2007], Tuzel et al. have introduced a modification to the original LogitBoost algorithm to specifically account for the Riemannian geometry. The modified LogitBoost algorithm is presented in figure 4.5

To train a level "k" of the cascade, a given number of weak classifiers are successively added. To add a weak classifier "l" to the current training classifier, 200 candidate weak classifiers are evaluated: 200 subwindows are randomly selected.

Let r_i be one of these subwindows and $\hat{\mathbf{C}}_{r_i}^j$ the corresponding normalized covariance descriptor on the sample j. For each subregion r_i, the weighted mean μ_i of all the normalized covariance descriptors $\hat{\mathbf{C}}_{r_i}^j$ of the positive samples is computed using a gradient descent procedure (eq. 4.20).

Using this mean μ_i, all $\hat{\mathbf{C}}_{r_i}^j$ of all the samples (positives and negatives) are projected onto the tangent space using (eq. 4.16) obtaining vectors in Euclidean space. Using these vectors and the corresponding weights of all samples, a regression fucntion g_i is computed.

The best weak classifier candidate, which minimizes negative binomial log-likelihood (eq. 4.23), is added to the current training classifier.

Input: Training set $\{(R_i, y_i)\}_{i=1...N}$, $y_i \in \{0, 1\}$

- Repeat for $k = 1...K$
  - Classify negative examples $\{R_i^-\}_{i=1...N_n}$ with the cascade of $(k-1)$ classifiers and remove samples which are correctly (negative) classified
  - Start with weights $w_i = 1/N$, $i = 1...N$, $F_k(R) = 0$ and $p_k(R_i) = \frac{1}{2}$, let $l = 1$
  - Repeat while $p_k(R_p) - p_k(R_n) < margin$
    - Compute the response values and weights $z_i = \frac{y_i - p_k(R_i)}{p_k(R_i)(1-p_k(R_i))}$, $w_i = p_k(R_i)(1 - p_k(R_i))$
    - Sample $\{r_{k,t}\}_{t=1...200}$ subwindows and construct normalized covariance descriptors $\mathbf{X}_{i,k,t} = \hat{\mathbf{C}}_{i,r_{k,t}}$
    - Repeat for $t = 1...200$
      - Compute the weighted mean of the positive samples $\{X^+_{i,k,t}\}_{i=1...N_p}$ through (24) $\boldsymbol{\mu}_{k,t} = \arg\min_{\mathbf{X} \in Sym_8^+} \sum_{i=1}^{N_p} w_i d^2(\mathbf{X}^+_{i,k,t}, \mathbf{X})$
      - Map the data points to the tangent space at $\boldsymbol{\mu}_{k,t}$ $\mathbf{x}_{i,k,t} = \mathrm{vec}_{\boldsymbol{\mu}_t}(\log_{\boldsymbol{\mu}_t}(\mathbf{X}_{i,k,t}))$
      - Fit function $g_{k,t}(\mathbf{x})$ by weighted least-squares regression of z_i to $\mathbf{x}_{i,k,t}$ using weights w_i
    - Update $F_k(R) \leftarrow F_k(R) + \frac{1}{2} f_{k,l}(R)$, where $f_{k,l}$ is the best classifier among $\{f_{k,t}\}_{t=1...200}$ which minimizes the negative binomial log-likelihood (29) and $p_k(R) \leftarrow \frac{e^{F_k(R)}}{e^{F_k(R)} + e^{-F_k(R)}}$
    - Sort positive and negative samples according to descending probabilities and find samples at the decision boundaries $R_p = (0.998N_p)$-th R^+, $R_n = (0.35N_n)$-th R^-
    - $l = l + 1$
  - Store $F_k = \{(r_{k,l}, \boldsymbol{\mu}_{k,l}, g_{k,l})\}_{l=1...L_k}$ and $thrd_k = F_k(R_n)$

Figure 4.5: Pedestrian detection with cascade of LogitBoost classifiers on Sym_8^+ (source [Tuzel 2007]).

The weights and the probabilities of all the samples are updated according to the new added weak classifier. The positive and the negative samples are sorted in a decreasing order using their probabilities. The current strong classifier is considered as fully trained if the difference between the probability of the $(99.8\%)_{\mathrm{th}}$ positive sample and the $(35\%)_{\mathrm{th}}$ negative sample is greater than 0.2. This means than the current trained cascade level has to reject at least 35% of the remaining negatives while it has to correctly detect at least 99.8% of the positive samples, and the value "0.2" represent a minimum margin to ensure between the 35% rejected negatives and the 99.8% positive. It is used to make the separation more reliable.

In this case, the training of the current cascade level is achieved. The negative sam-

ples are tested with the new cascade and all correctly classified samples (recognized as negatives) are removed from the training dataset. The next cascade level is trained using remaining negatives.

The training algorithm produces a set of K LogitBoost strong classifiers.

Yao et. al [Yao 2008] have introduced two important improvements that we use also.

First, classifiers are trained on a covariance descriptors with lower dimension. The initial set of 8-dimensional features is always kept and all the corresponding Integral Images (for individual features and the feature products) are computed, but a weak classifier is defined by a lower dimensional covariance matrix, computed with the more representative subset of features in the considered image subregion, and can vary from a region to another one.

Practically, each candidate subregion among the 200 randomly selected does not provide a unique candidate weak classifier with a 8×8 covariance matrix, but a set of candidate weak classifiers, each of them is computed using a 4×4 covariance matrix. All the possible combinations of 4 features from the 8 initial ones are considered.

This provides $\binom{8}{4} = 70$ covariance matrices to consider for each candidate subregion. To avoid the high computational cost required to test all these 70 covariance matrices for each subregion, a substitute of the negative binomial log-likelihood for each 4×4 matrix for each subregion is calculated as follow:

- The $\binom{8}{2} = 28$ possible 2×2 covariance matrices of the all possible pairs of features (from the 8 initial ones) are computed for each subregion.
- The negative binomial log-likelihood for each 2×2 matrix is computed using (eq. 4.23).
- Instead of computing the real negative binomial log-likelihood of a given 4×4 matrix computed with the four features $\{f_1, f_2, f_3, f_4\}$, this value is replaced by the sum of the negative binomial log-likelihood of all the 2×2 matrices which are computed with at least one of $\{f_1, f_2, f_3, f_4\}$ features.
- The subset of 4 features which provides the minimal sum is considered as the best (more representative) subset for the considered region.

The second improvement consists in the concatenation of the mean feature vector of each random subregion to its corresponding mapped vector of each sample before regression computing, improving performances by increasing the amount of information in the final vector.

4.3.3 Cascade of Classifiers Optimization

4.3.3.1 Main Issues

The structure of the cascade of classifiers is as important as its content. The cascade content (weak and strong classifier discriminative power) defines the detection performances in terms of false positive (false detection) and false negative (miss detection) rates, while the cascade structure (the number of strong classifiers in the cascade and the number of weak classifiers in each of them) greatly defines the processing time cost.

The discriminative power of a weak classifier depends directly on its selected subregion (the information contained in this subregion). The main issue in the proposed method is related to the fact that each weak classifier is selected as the best one from "n" randomly selected candidate weak classifiers (n=200 is taken in the approach) but not the best possible one at all. For the iteration "l", the best weak-classifier candidate from the "n" randomly selected ones can be insufficiently discriminant, and its selection will require to add more weak-classifiers to compensate its low discriminative power, and thereby lengthening the strong classifiers of each cascade level.

The optimal cascade of classifier in term of discriminative power and time processing can be obtained by selecting the best weak classifier from all the possible ones, at each iteration "l" of each level "k", because in this case, we ensure that the selected weak classifier is the most discriminative one, but in practice, this is impossible to perform (or at least, in a reasonable training time).

In fact, if we take the INRIA training dataset as example, it contains people training images with 64×128 pixel size in which, the person image is surrounded by a 16 pixel margin. Considering the minimum size of subregions as 10% of image width and height as proposed in [Tuzel 2007], the number of all possible subregions with a minimum size of 6×12 pixels is 12.218.310, and if only person sub-image is considered (ignoring the margin), i.e. images of 32×96 with subregions of 3×9 minimum size, the number of all possible subregions is 1.820.940 subregions.

This represent a high number of subregions to process at each iteration "l" of each level "k", knowing that even if it is possible to compute the normalized covariances of all these subregions on all training images once before training (because the covariance matrices do not change), the mean covariance of each subregion will change from training iteration to another one, due to the used weighting process which evolves according to the previous iteration and to the considered samples at a given time. This requires to compute a weighted mean of more than 1.8 million covariance matrices (or more than 12 million covariances matrices in the case of the whole images, with margin), using the gradient descent procedure (eq. 4.20) which is an iterative procedure, requiring several

iterations. In addition to this, the minimal representation of each covariance matrix in tangent space and the regression function will change due to the dependency of the projection function to the mean covariance matrix.

By using only 200 subregions at each iteration, the training time indicated in [Tuzel 2007] is approximatively 2 days. We can easily see that processing more than 1.8 million subregions (or 12 millions subregions depending on the inclusion of margin) is not reasonable. It is then necessary to select a subset of subregions at each iteration, but without any a priori knowledge concerning the more informative regions, this selection is performed randomly, increasing the probability to have more weak classifier in each cascade level than in the optimal case.

The previously described people detector provides interesting detection performances, with a lower rate of miss-detections and false positives (see 7.1), but it has the disadvantage of being highly time consuming for the detection process and not applicable for real-time processing. In [Tuzel 2007], the authors indicate that detection time on a 320 x 240 image is approximatively 3 seconds for a dense scan, with 3 pixel jumps vertically and horizontally.

The feature subset selection approach, proposed by Yao and Odobez in [Yao 2008] allows to work in a lower dimensional symmetric positive definite matrices, making eigenvalue decomposition faster and thereby allowing real time processing, but the processing time reduction is greatly due to the use of foreground features, extracted using background subtraction algorithm. Non moving image regions are quickly rejected. Unfortunately, the unavailability of high quality foreground information (as the ones used in [Yao 2008], shown in figure 4.3), and some time the unavailability of the background subtraction (in static image case) limits this improvement dedicated to video sequences only, with the availability of an efficient background subtraction.

Note that most computationally expensive operation during the training and the classification is eigenvalue decomposition. This decomposition is the basis of all operators in Sym_d^+. Eigenvalue decomposition of a symmetric $d \times d$ matrix requires $O(d^3)$ arithmetic operations.

We focus in our work on another way to make the classification faster while maintaining high classification performance. At the end, the obtained approach improves also the training stage. This is performed by compensating the random selection of subregions issue with a pre-training step.

In the following paragraphs, we will illustrate the improvements of our contribution using tests on INRIA Person dataset [], due to the availability of more information concerning cascade structure and processing time in [Tuzel 2007] on this dataset, but the reasoning and the proposed method can be generalized to all datasets.

4.3.3.2 Clustering Negative Data Before Training

Using a large number of samples slows down the training process. Of course, the larger the training dataset is, the more efficient the classifier cascade is. But most of the time, especially for first cascade levels, a large number of negative samples contains very similar information. This is due to the fact that a new level is trained using false positives of previous levels, and in this case, these false positives are generally resulting from successive small shifts of testing window on the image, providing very similar content.

The trained cascade of classifiers on INRIA Person dataset by Tuzel et al. in [Tuzel 2007] contains 30 strong classifiers (levels) (see figure 4.6 (a)).

A first idea to speed up the training process is to use a smaller subset of randomly selected negative samples to train a given cascade level. We can suppose that a randomly selected subset can be statistically representative of all remaining negatives, and a trained cascade level on this subset will reject a proportional number of negatives from the whole training dataset than the one rejected from the selected subset.

We have tested this approach and we have observed that random selection effectively speeds up slightly training and provides 4 less cascade levels in comparison than [Tuzel 2007] but with longer classifiers (figure 4.6 (b)), slowing down the detection in comparison to the previous mentioned approaches. In fact, one cascade level consists of a set of weak classifiers. The response of one classifier is obtained after computing the output values of all its weak classifiers. It means than a long classifier containing a large number of weak classifiers takes more time to return a decision, so a cascade of long classifiers is very slow for detection.

Decreasing the training time by increasing the detection one is not acceptable as long as the detection process is the final aim, but the idea of training each cascade level on a subset of negative samples to speed up the training is still an interesting way to explore. It is just necessary to find the best way to select the negative sample subsets to decrease the number of necessary weak classifiers per cascade level and to take into account the random nature of the subregion selection.

We have constructed our reasoning starting from this observation: the number of weak classifiers per cascade level depends mainly on the diversity of negative samples used for the training. The characterisation of positive samples and their separation from the negative ones require as many subregions of interest as the samples are diverse.

To illustrate the relationship between negative sample diversity and classifier cascade structure, let us use a simple example which can be generalized to understand the concept.

Suppose that we have to separate a person image from three non-person images: a

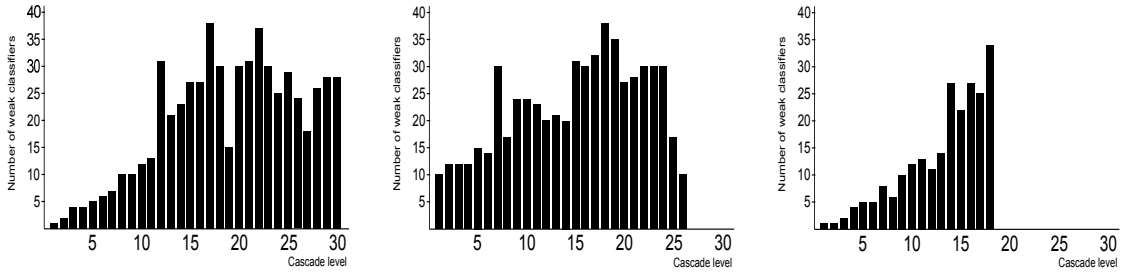


Figure 4.6: Comparison between structures of the cascade of classifiers: (a) Tuzel et al.[Tuzel 2007]. (b) Random subset selection: shorter cascade (4 levels less than original approach) but more weak classifiers in most of cascade levels. (c) our proposed approach: less cascade levels with less weak classifiers in most of them in comparison with (a) and (b).

blue sky image, a vertical barrier image and a lamppost image.

- Due to the poverty of texture and gradient on the blue sky image, a unique large covariance region (figure 4.7(a)) is sufficient to separate the blue sky image from the person image, which has many gradients and a vertical shape.
- For vertical barriers, the previous region is not appropriate due to the vertical shape of a barrier. A smaller region around the person's head is more appropriate (figure 4.7(b)). The circular shape of the head provides a good separation between a person and a vertical barrier.
- Now, for the lamppost, the two previous regions are not suitable. It is necessary to take a region around legs to encode the separation between the legs (figure 4.7(c)).

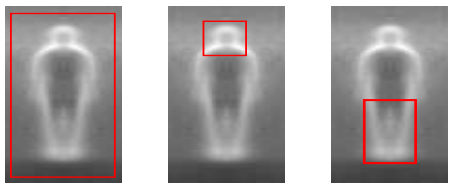


Figure 4.7: Three possible weak classifiers to reject: (a) low texture and non vertical shapes. (b) non circular shape at the top (head). (c) not separated shape at the bottom (legs).

For this example, there are two methods to train the classifier. The first cascade is trained with the three negative images at the same time, using appropriate parameters. The second cascade is trained with one negative image at a time in the mentioned order.

The first method provides "a cascade" of a unique strong classifier containing three weak classifiers at least (the number can be larger due to the possible combinations) (figure 4.8(a)). The second method provides a cascade with three levels (figure 4.8(b)), each level containing one weak classifier corresponding to one case (textured and vertical shapes only, circular shape at the top of vertical shape, separation between legs).

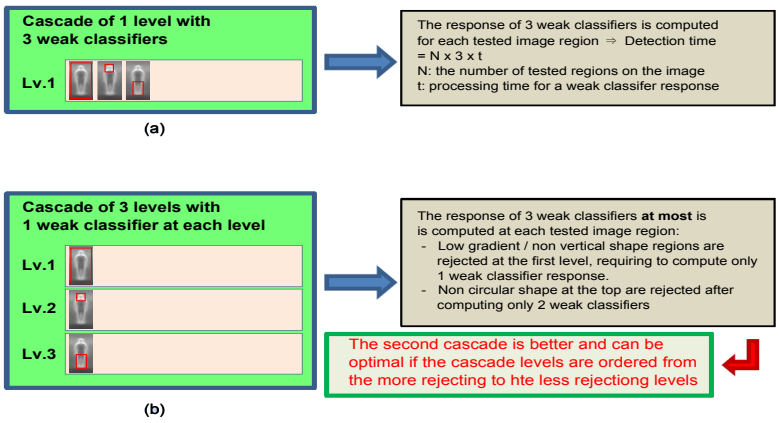


Figure 4.8: Two possible cascade structures: (a) One level cascade with three weak classifiers. (b) Three levels cascade, with one weak classifier per level. The second cascade is more optimized.

Suppose now that we have to perform a people detection on a large image which contains only sky, a low textured road, some vertical barriers and some lampposts. Both cascades will provide equivalent detection performances, but the second one will be faster. This is because most of tested windows (sky and road) are rejected after evaluating only one covariance descriptor (the one of the first cascade level), while the second classifier cascade needs to evaluate three (or more) covariance descriptors for each tested window.

We propose an approach using a smaller subset of negatives at each cascade level training to make it faster. Our approach provides shorter cascade with smaller classifiers on average (figure 4.6 (c)) in comparison with the [Tuzel 2007] one (figure 4.6 (a)) speeding up the detection process. At the same time, the experimental results show that our approach provides slightly better detection performance than the original one.

The idea consists in regrouping negative samples per groups containing similar content in terms of covariance information, and in training each cascade level with one group of similar samples.

The previously described Logitboost algorithm achieves characterization of people against a group of negative samples faster when these negative samples are more similar.

It also specializes each cascade level faster and reduces the effect of random subregion selection method for best weak classifier extraction in this case.

We have tested two clustering methods to achieve the negative sample regrouping. The first one is performed directly in the Riemannian manifold of covariance matrices while the second one is performed in the Tangent vector space.

4.3.3.3 Hierarchical Clustering in Riemannian Manifold of Covariance Matrices

The first clustering method to group similar negative samples in term of covariance information is performed directly using the distance between covariance matrices provided by eq. (4.15).

To compare two negative images in term of covariance information, and due to the unavailability of a priori knowledge concerning the most important subregions in these images, we perform a pyramidal division of the images in 4 levels where level 0 is the whole image (see figure 4.9).

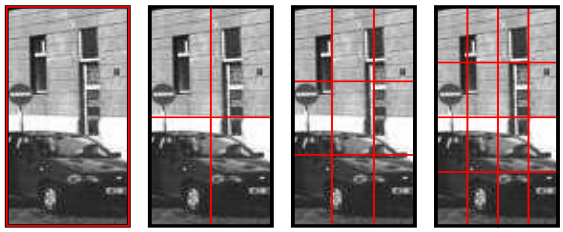


Figure 4.9: The 4 levels pyramidal subdivision of negative images for clustering. The level 0 (the left one) is the whole image.

The distances between all pairs of negative images is performed in each pyramid level. For a given pyramid level, the distance between two negative images I_1 and I_2 is given by the sum of squared distances between all pairs of subregions at the same location from the two images. The distance between two image subregions is provided by the distance between their corresponding covariance matrices using eq. (4.15) (see figure 4.10).

A triangular matrix containing distances between all pairs of negative images is then computed for each pyramid level. From a given matrix of distances, it is easy to extract iteratively the largest cluster which contains the most similar n% of images. We take the value of 35% of remaining negative samples to be similar and comparable to [Tuzel 2007] and [Yao 2008]. The hierarchical clustering is illustrated in figure 4.11.

From each pyramid level, we can extract the largest clusters of the most similar remaining negative samples. This provides 4 candidate largest clusters, the largest one from each pyramid level. We select the best cluster as the one which provides the lowest

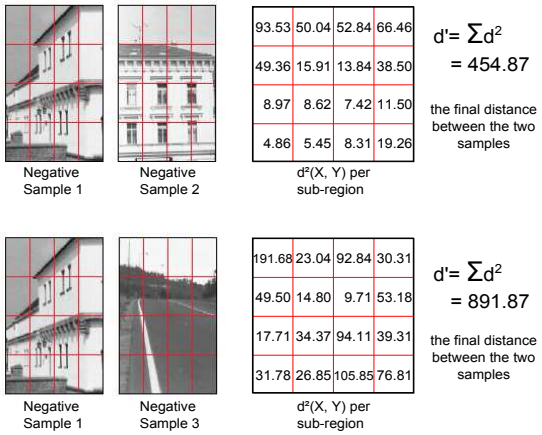


Figure 4.10: Negative image distance in the last pyramid level (4×4 subdivision), computed as the sum of squared distances between each pair of covariance matrices of the same subregion.

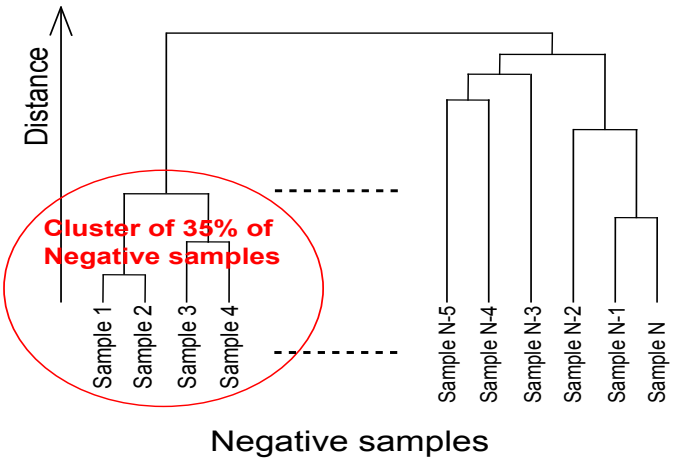


Figure 4.11: Hierarchical tree of clustered negative samples.

inter-images distances. This value is computed as the sum of all distances between all pairs of images belonging to the considered cluster.

The training is now done by using the largest cluster of most similar negative images for each trained cascade level. Once a cascade level is trained using the selected cluster, the new cascade is applied to all the remaining negative samples, those used for training and the others which are not in the selected cluster. The next cluster is selected using the same method described below, applied on the remaining negative images.

Note that for a given cascade level, we observe that 80% to 95% of the negatives from the used cluster are correctly classified and removed and a small part of unused negative images (not belonging to the used cluster) also.

The comparison between the structure of our cascade of classifiers and the one from [Tuzel 2007], shown in figure 4.6, shows the effectiveness of our method. Our cascade of classifiers is shorter than [Tuzel 2007] one and most levels in our cascade of classifiers are shorter than their corresponding levels in [Tuzel 2007] one.

A more detailed comparison in terms of detection performances and processing time and a global evaluation which validates the effectiveness of the proposed people detector are provided in the chapter 7.

The second tested clustering approach consists in projecting all the remaining negative samples to the tangent space on their mean points. In this method, we consider the covariance matrix of the whole image as its representation (equivalent to the level 0 of the used pyramid in the previous method). The mean of all negative samples is computed and used to project all covariance descriptors to the Euclidean space. Finally in Euclidean space, the clustering is performed using adaptive bandwidth mean shift filtering [Comaniciu 2002]. (See Figure. 4.12)

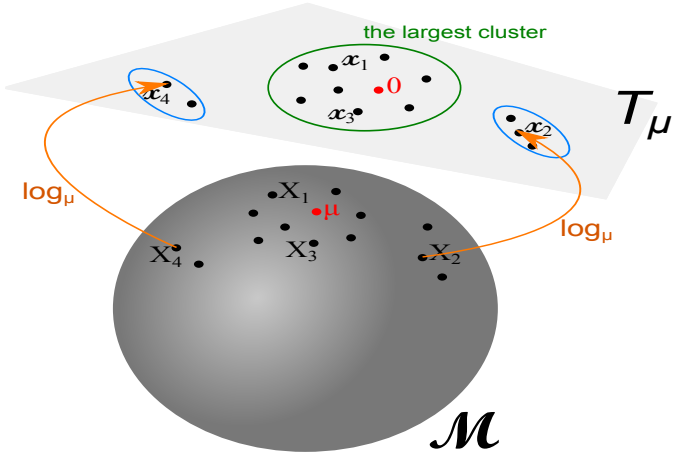


Figure 4.12: Illustration of clustering on a tangent space to a 2D Manifold.

The obtained cascade of classifiers, trained using clustered negative samples by this second method is substantially similar to the one we have obtained with the original method (without clustering). This is due to the negative sample sparsity.

Unlike for positive samples, the mean of negative samples does not have a physical sense (see figure 4.13, second row), or at least, it has a sense for the largest clusters of the first training levels which share some similar content. Positive samples share a similar shapes, even if variations can occur in these shapes (see figure 4.13, first row). The mean covariance matrix of a given subregion can be considered as the covariance matrix of the mean shape of this subregion on all positive samples.

Generally, remaining negative samples after few iteration do not share similar shapes

in the same image locations, so the resulting mean does not represent a physical information from the considered negative samples. Therefore, projecting negative samples using insignificant mean loses the topological distribution of these negative samples in the tangent space. This is illustrated in figure 4.14

For this reason, our final people detector is trained using the hierarchical clustering in Riemannian manifold of covariance matrices.

Figure 4.13: Mean of gradient images. First row: some various positive image samples with the mean gradient image of the positive training dataset. Second row: some various negative image samples with the mean gradient image of the negative training dataset. The mean of positive images represent a mean shape of the human body while the mean of negative images does not represent any shape.

4.4 Conclusion

We have proposed an approach to optimize people detection using covariance descriptors. This approach consists in clustering negative data before the training step to obtain better classifier structure. The resulting detector is faster than original one and was trained in shorter time (detailed evaluations are provided in chapter 7).

Clustering negative data before training allows to reduce the effect of random subregions selection for best weak classifier training. As explained above, the exhaustive testing of all possible weak classifiers at each iteration is impossible in a reasonable processing time, and a targeted subregion selection is not possible due to the unavailability of any a priori information concerning the most interesting regions.

Of course, the unavailability of any a priori information concerning the most interesting regions can be considered as an issue for the negative image comparison and clustering. To compare two negative images and compute a distance between them, it is necessary to focus on the region of interest on each of them, which is not possible

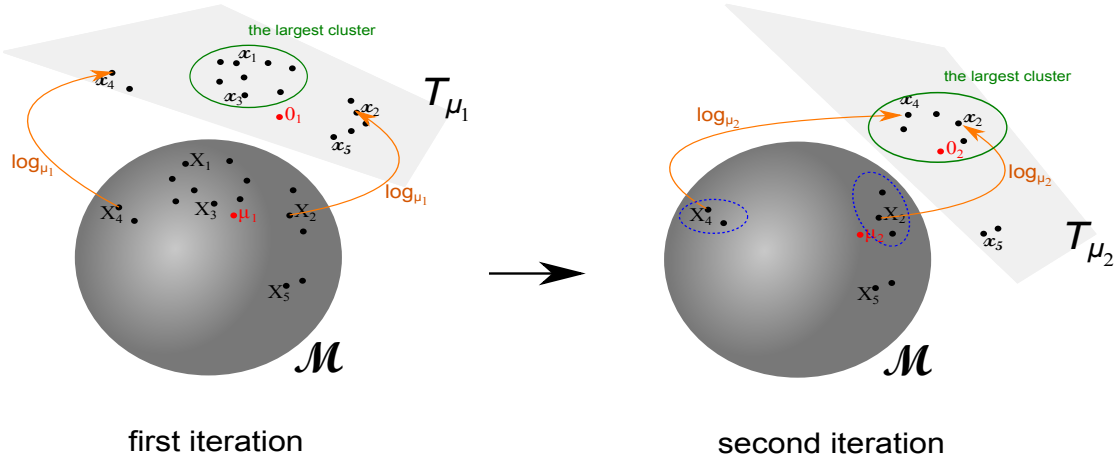


Figure 4.14: Illustration of negative samples sparsity in a 2D Manifold: In the first iteration, many negative samples are available. The largest cluster in the Riemanian manifold $\mathcal{M}$ attracts the mean near to it, and the projected points keep a large amount of topological distribution. Thereby, the first cluster has a sense in term of covariance matrices similarity. In the second iteration, once the largest cluster is removed, the negative samples are sparse, and the computed mean provides a projection with a changed topological distribution. The samples X_2 and X_4 which are clearly different in $\mathcal{M}$ are grouped in the same cluster (x_2 and x_4).

for the same reason. To deal with this, a clustering stage is performed using several subdivisions of images. It is not the optimal way but it reduces significantly the random nature of subregion selection during the training.

We have shown that this clustering is better when it is performed directly in Riemannian manifold of covariance matrices, due to the fact that a distance between two covariance matrices is more precisely represented by the length of the geodesic which links them, instead of the euclidean distance between their projected points in the tangent space. This euclidean distance may be altered by a non-significant mean computed on sparse negative data.

The basic method [Tuzel 2007] on which we have built our reasoning and improvements are not the last and the most efficient ones in the state of the art [Dollar 2010, Walk 2010, Felzenszwalb 2010]) when this thesis has been written, but when we have performed this work, [Tuzel 2007] was one of the most efficient approach in the state of the art , and our interest for this method is also justified by the nature of the trained classifier and the high processing time required. It allows us to highlight the contribution of our proposed method and to discuss its generalisation to other approaches based on cascade of classifier training.

In fact, clustering negative data before training is an efficient way to optimize the trained cascade of classifiers, for which the training cannot explore all the space of

possible weak classifiers to select the best one at each iteration. According to the used descriptor (Haar-like, LPB, etc.), to its similarity measure, and to the way the candidate weak classifiers are preselected for the selection of the best one, it is possible to speed up the training phase using this clustering method in the appropriate descriptor space (generally in Euclidean space for most of descriptors).

5

ROBUST OBJECT TRACKING USING PARTICLE FILTERING

This chapter describes the proposed object tracking algorithm in mono-camera context. This tracking is performed with static and calibrated cameras. The tracked objects are modelled with a set of SIFT features, selected in a specific way. The object tracking is performed in three separate levels: first the SIFT features are tracked independently using a particle filter. Then, object localisation and temporal links are built using a data association framework based on the localisation and the reliability of the tracked SIFT features. Finally, occluded objects are managed with two methods, the first one, based on real world information and dominant color descriptor, is faster and is sufficient to deal with occlusion situations in most of the cases, but for ambiguous or complex cases, the re-identification method, presented in chapter 6 is used.

5.1 Tracked Target Initialization

In the proposed tracking algorithm, moving object detection in the scene is performed using a state of the art background subtraction algorithms based on adaptive background mixture models [Stauffer 1999]. In this algorithm, moving image regions are extracted as group of foreground pixels. These groups of pixels are more or less connected, depending on the contrast of the corresponding object and some parameters of the background subtraction algorithms (several thresholds, number of considered Gaussian, etc.). A clustering step is performed to regroup the foreground pixels in blobs representing one or several grouped objects in the scene, and to filter too small groups

of pixels which are probably noise (see figure 5.1). Finally, these blobs are delimited by minimal bounding boxes (the smallest ones which surround blobs) which will be used to define the localisation of the objects in image.

Figure 5.1: Background subtraction results for real video. The top-left image is the original one, the three other images are the results of background subtraction using different parameter values [McHugh 2009]

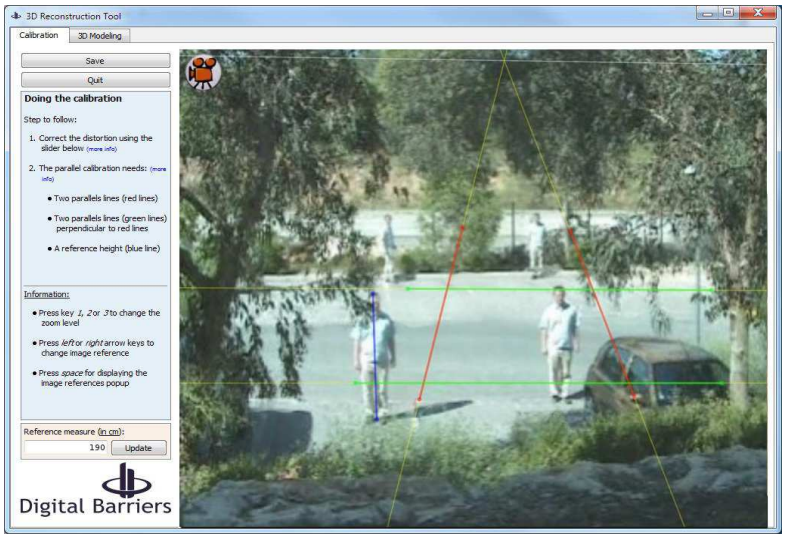


Figure 5.2: Screenshot of Digital Barriers calibration tool

5.1.1 Classification by Real Dimension Estimation

As mentioned in the introduction chapter, our work is performed using calibrated cameras. The camera calibration is performed using a proprietary software developed

by Digital Barriers France (previously Keeneo). The calibration is based on vanishing lines approach (see figure 5.2). It provides both extrinsic and intrinsic calibration matrices in an Y-top coordinate system. These matrices are used for **real world ⇄ image projection**. Note that this tool is easy to use and requires few operations from video operators (defining 4 straight lines on the ground floor and one vertical line, in addition to an estimation of the height of any object in the scene). This does not affect the "easy to use" constraint we try to respect.

A first classification of the moving objects is performed using an estimation of their real dimensions (3D width "W" and 3D height "H") provided by camera calibration matrix (see figure 5.3).

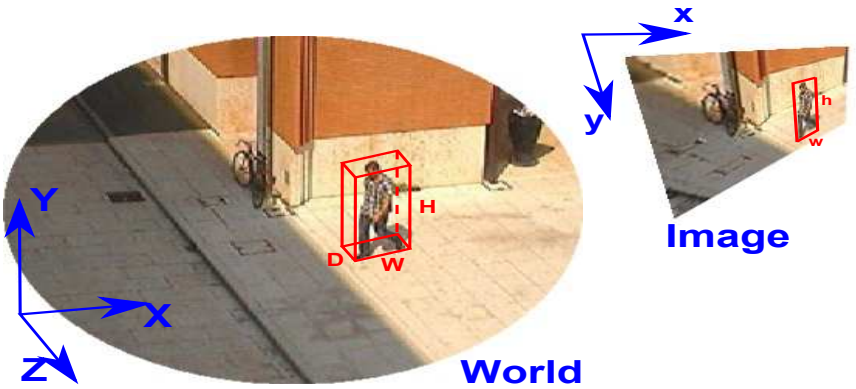


Figure 5.3: Real world dimension projection. Depth is ignored

Any real world point $P_1(X, Y, Z)$ has a unique projection $p_1(x, y)$ in the image, but a given 2D image point $p_2(x, y)$ has infinite possible corresponding points in real world. All these points belong to a straight line (L) which passes through the optical center of the camera (see figure 5.4). To obtain a unique corresponding point (X, Y, Z) in the real world for a 2D point (x, y) in the image, an additional constraint is required. Generally, the ground plane is considered as the plane with coordinate $Y = 0$, then all 2D points on the ground image can be easily projected to real world and their real coordinates in the camera coordinate system can be computed.

- To estimate the real width W of a given object in the scene, the two bottom corners of its bounding box p_{bl} (bottom-left point) and p_{br} (bottom-right point) are projected with $Y = 0$ assuming that the object is on the ground. Once the two real coordinates of the object borders P_{bl} and P_{br} are computed, the distance between them is considered as the object width W (see figure 5.5).

- For the real height H estimation, a double constraint is considered. The bottom-center point of the bounding box p_{bc} is projected to real world coordinates P_{bc}

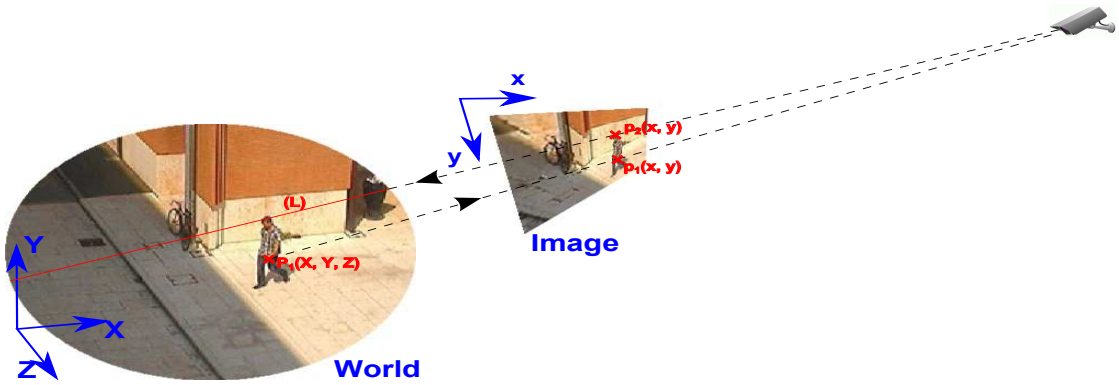


Figure 5.4: World to image and image to world projections: a unique image point $p_1(x, y)$ corresponds to a given real world point $P_1(X, Y, Z)$ but an infinite number of real world points, belonging to the straight line (L) correspond to a given image point $p_2(x, y)$

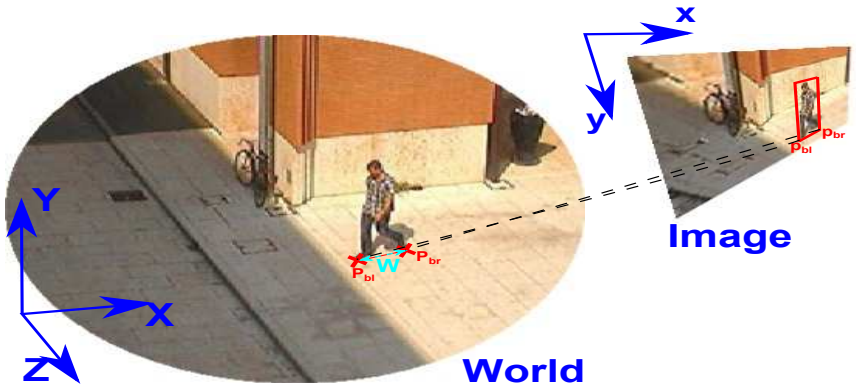


Figure 5.5: Real width W estimation

using $Y = 0$ providing a first point $P_{bc}(X_1, 0, Z_1)$. The top-center point of the bounding box is also projected in real world using $Y = 0$, providing second point $P_2(X_2, 0, Z_2)$ but this point does not really represent the top point of the object. It represents the intersection point between the ground plane (G) on one hand and the straight line (L) passing through the optical center of the camera and the top point of the considered object in the other hand. To estimate the real coordinates of the top point of the object, the intersection point between this straight line (L) and the plane (π) containing the two points P_{bl} and P_{br} and which is perpendicular to the ground plane (G) is computed and considered as the wanted point P_{tc} (real top-center point of the object). The distance between P_{bc} and P_{tc} is considered as the object height (see figure 5.6).

The depth of objects is ignored due to the impossibility to estimate it from a single

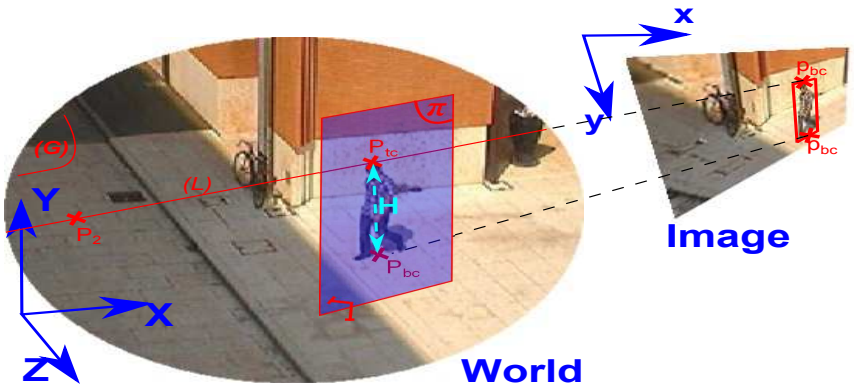


Figure 5.6: Real height H estimation

monocular camera.

The first classification of moving objects is then performed by fitting the real dimensions with previously defined dimension ranges for known classes of objects of interest (see figure 5.7).

5.1.2 People Detection

The previously detailed method is fast to perform and provides, most of the time, good detected and classified objects, but can be imprecise in some situations. For example, if no shadow removal is performed, the bounding box representing the object is not precise and includes shadow pixels, providing erroneous projections and real dimensions which did not fit with known classes or fit with incorrect ones. Another example concerns grouped objects. If a group of individuals are too close to each others, the extracted bounding box will include all the people and its projected dimensions will not fit with any class or will fit with wrong one (see figure 5.8).

To deal with these possible situations for people tracking, while keeping the interesting aspect of fast detection and classification using background subtraction and camera calibration information, we use our people detector following two rules:

- The people detector is applied on the moving regions on which the first classification is ambiguous (object dimensions which do not fit with any class dimensions, large objects which can be a group of people, etc.). This allows to reduce the area of searching to a subregion instead of the whole image.

- The calibration information is used to limit the searching scales of the people detector. The admitted interval of people height (the human model) is used to project

realistic searching window size on the moving region, avoiding useless scale detection.

These two rules allow to perform the detection in negligeable time, and thereby, did not slow all the tracking process which is still performed in real time.

Note that the real world dimension estimation and the background subtraction information are also used during the tracking itself. One for fast occlusion management and the other for the particle filtering process. This will be detailed in the following sections.

Our tracking algorithm requires the set of all detected moving objects of interest in every frame of the video sequence, in the form of bounding boxes and their corresponding extracted foreground pixels.

5.2 Object modeling

To be tracked, an object of interest has to be modeled using its discriminant features. Among the different possible representations for an object (colors, shapes, etc.), we choose to model each tracked object by a set of features points. This choice is motivated by the independence of points between them, which allows to deal with partial occlusion and object deformations.

Our tracking algorithm is built on SIFT features for object modeling, but any other kind of point of interest around which a descriptor can be computed (SURF [Bay 2008], HOG [Dalal 2005], etc.) and can be used with slightly lesser point tracking performances. SIFT features are known to be among the most robust local descriptors. Even if some authors have claimed that their proposed new descriptors, like SURF [Bay 2008], are faster than SIFT and provide comparable performances, these affirmations were probably correct since years ago but actually, using recent computers (more processing power) and optimized implementations, we have observed than no significant difference in processing time exists between SIFT and the other similar local descriptors.

5.2.1 SIFT features

Scale-Invariant Feature Transform (SIFT) is an algorithm in computer vision to detect and describe local features in images. It was first proposed by Lowe [Lowe 1999]. SIFT features correspond to a set of points of interest called SIFT keypoints which are detected in a specific way, and around which a local descriptor, called SIFT descriptor, is computed and assigned to them.

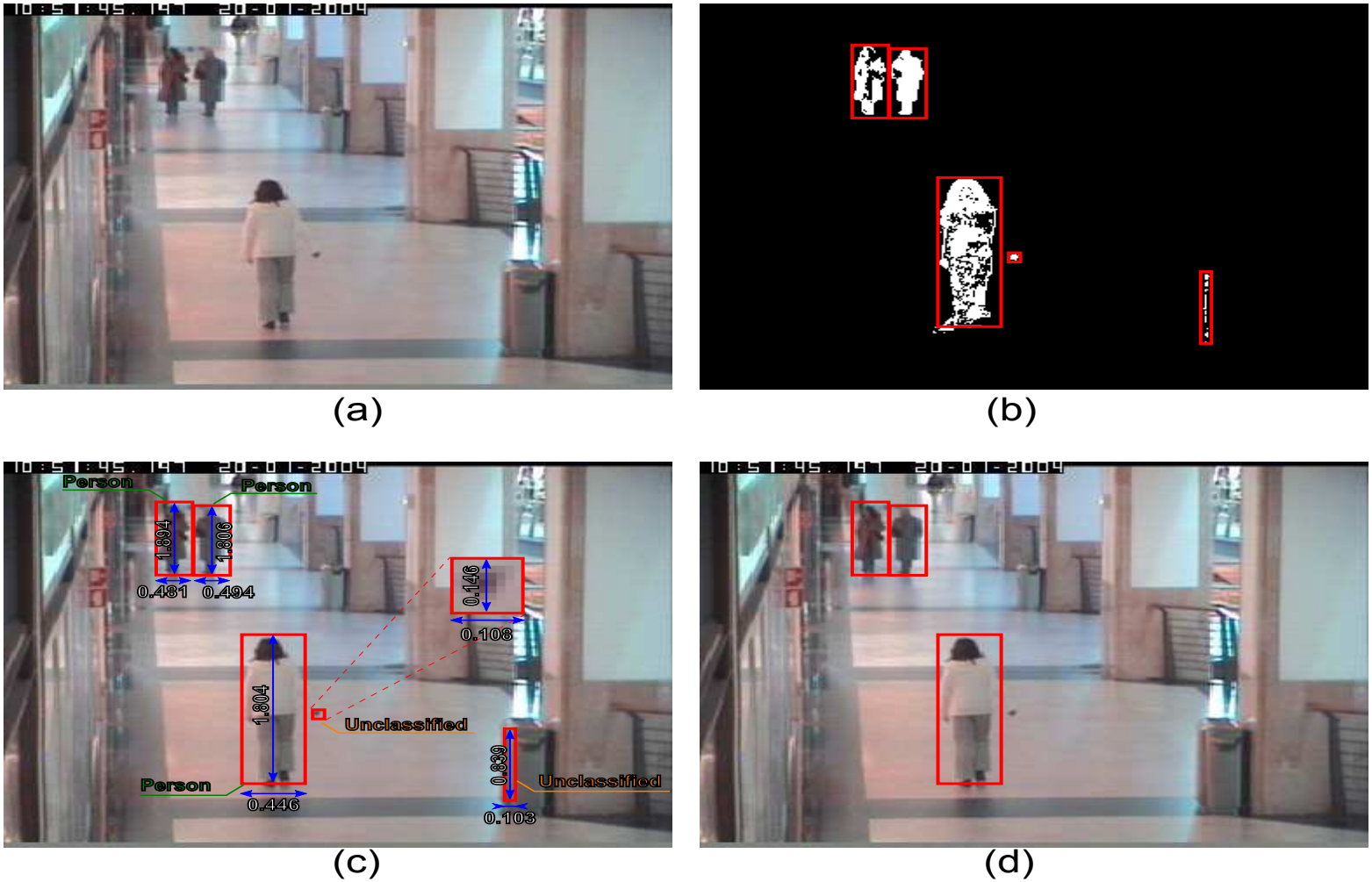


Figure 5.7: Object classification by fitting estimated real dimension in known model dimensions. (a) Original frame. (b) Background subtraction results and boundingbox creation after clustering. (c) Estimation of real dimensions of delimited objects (in meter) using camera calibration. (d) Filtering of unclassed objects which did not fit with known models and validation of objects of interest (people)

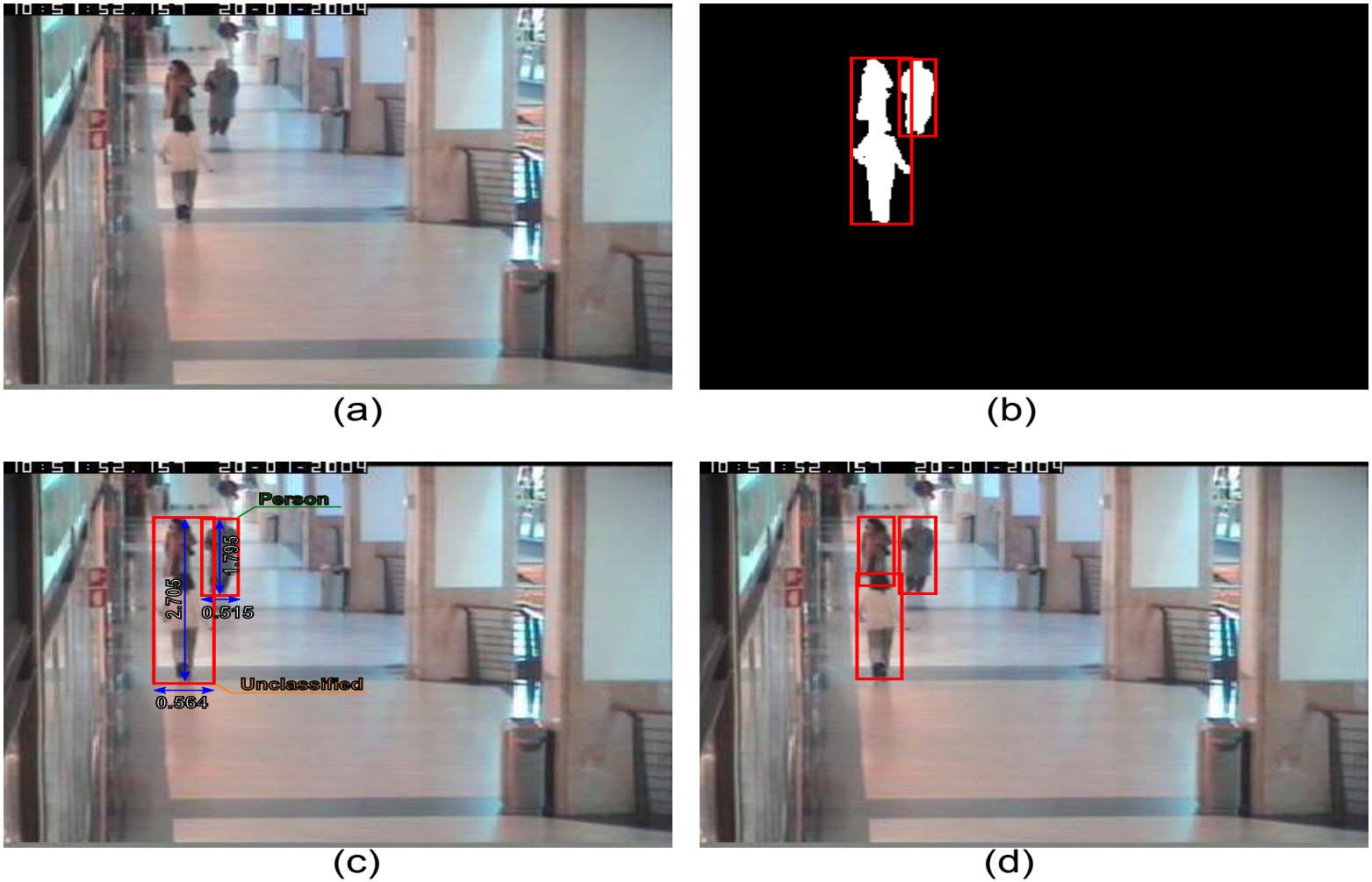


Figure 5.8: Use of people detector to split grouped persons. (a) Original frame. (b) Background subtraction result. Two persons are grouped due to the connected foreground pixels. (c) Failure of the classification of grouped people by real dimension estimation. (d) Result of our people detector applied in the restricted area of unclassified object (by real dimension estimation)

5.2.1.1 SIFT Points

The SIFT point detection is performed in four successive steps. Each step provides to these points some robustness against a kind of issues or variations.

- **Scale-space extrema detection**: A pyramid of multiple octaves containing Difference of Gaussian images is built (see figure 5.9 (a)). The image is convolved with Gaussian filters at different scales $k\sigma$, and then the difference of successive Gaussian-blurred (DoG) ($D(x, y, \sigma)$) images are taken. DoG images are given by:

$$D(x, y, \sigma_i) = L(x, y, \sigma_{i+1}) - L(x, y, \sigma_i) \tag{5.1}$$

where

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{5.2}$$

$I(x, y)$ is the original image and $G(x, y, k\sigma)$ is a Gaussian filer at scale σ.

Candidate SIFT points are then taken as extrema of the Difference of Gaussians (DoG) that occur at multiple scales. Each pixel in the DoG images is compared with its eight neighbors at the same scale and nine corresponding neighboring pixels in each of the neighboring scales (see figure 5.9 (a)). If the pixel value is the maximum or minimum among all compared pixels, it is selected as a candidate keypoint.

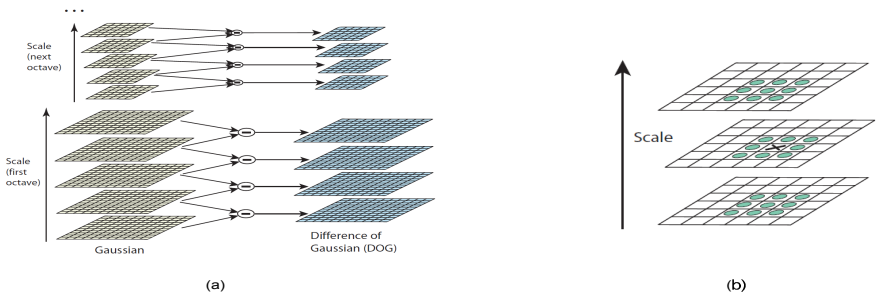


Figure 5.9: Extrema points detection on the DoG (Difference of Gaussian) pyramid. (a) The build multi-octave pyramid. (b) Extrema detection (source [Lowe 2004])

- **Keypoint accurate localisation**: The selected extrema are (re)localised more accurately by interpolating their initial positions using the quadratic Taylor expansion of the Difference-of-Gaussian scale-space function, $D(x, y, \sigma)$ with the candidate keypoint as the origin. This Taylor expansion is given by:

$$D(\mathbf{x}) = D + \frac{\partial D^T}{\partial \mathbf{x}}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x} \tag{5.3}$$

where D and its derivatives are evaluated at the candidate keypoint and $\mathbf{x} = (x, y, \sigma)$ is the offset from this point. If the offset is larger than 0.5 in any dimension, then that is an indication that the extremum lies closer to another candidate keypoint. In this case, the candidate keypoint is changed and the interpolation performed instead about that point. Otherwise the offset is added to its candidate keypoint to get the interpolated estimate for the location of the extremum.

- **Low-contrast keypoint elimination**: The value of the second-order Taylor expansion (eq. 5.3) is computed at the new localisation of each keypoint. If this value is less that 0.03 (this contrast threshold has been fixed in [Lowe 2004] as the optimal one), it is considered as a low contrast point and it is eliminated.

- **Edge response elimination**: The keypoints that have poorly determined locations (during the keypoint accurate localisation step) but have high edge responses are eliminated. This is performed using the second-order Hessian matrix H defined as:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{5.4}$$

The eigenvalues of $\mathbf{H}$ are proportional to the principal curvatures of D. The curvature value is taken as $r = \alpha/\beta$, where α is the larger eigenvalue of H and β its smaller one. The trace of H ($D_{xx} + D_{yy}$) gives us the sum of the two eigenvalues of H, while its determinant ($D_{xx}D_{yy} - D_{xy}^2$) yields the product. The ratio $R = \mathrm{Tr}(\mathbf{H})^2 / \mathrm{Det}(\mathbf{H})$ can be shown to be equal to $(r+1)^2/r$, which depends only on the ratio of the eigenvalues rather than their individual values. R is minimum when the eigenvalues are equal to each other. Therefore the higher the absolute difference between the two eigenvalues, which is equivalent to a higher absolute difference between the two principal curvatures of D, the higher the value of R. It follows that, for some threshold eigenvalue ratios r_{th}, if R for a candidate keypoint is larger than $(r_{th}+1)^2/r_{th}$, that keypoint is poorly localised and hence eliminated. In [Lowe 2004], $r_{th} = 10$ is taken as the optimal value.

The remaining candidate points are the final SIFT points. These steps provide invariance to image location and scale to the detected SIFT points. They also ensure a sufficient contrast and avoid curvature to them.

5.2.1.2 SIFT Descriptors

Once SIFT points are detected, the next step consists in computing SIFT descriptors around each of them. This is performed in two steps:

- **Main orientation assignment**: To achieve invariance of SIFT descriptor to 2D rotations (in image), each keypoint is assigned with one or more orientations based on local image gradient directions. For a SIFT point (x, y, σ), the nearest Gaussian image $L(x, y, \sigma')$ to the DoG image on which this point is detected is taken to construct an orientation histogram of 36 bins (each bin covering 10 degrees). Each pixel in a neighboring window of the considered SIFT point will contribute to this histogram by its weighted gradient magnitude. The weighting is provided using a circular Gaussian windows centred on the SIFT point. The gradient magnitude and orientation of each neighboring pixel are computed as:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (5.5)$$

$$\theta(x, y) = \text{atan2}(L(x, y+1) - L(x, y-1), L(x+1, y) - L(x-1, y)) \quad (5.6)$$

  Once the histogram of 36 bins is computed, the orientation corresponding to the highest peak is taken as the main orientation and assigned to the SIFT point. If any other local peak value is greater than or equal to 80% of the highest peak value, the corresponding orientation is also taken as another main orientation. In this case, the SIFT point is duplicated and each of these SIFT points will be assigned with one main orientation. There will be as many SIFT points at the same location and same scale as the number of main orientations.

- **Descriptor computing**: For each SIFT point, the image closest in scale to its own scale is taken for computing the descriptor. On this image, a square region of 16×16 around the SIFT point is considered. This window is divided in 4×4 equal subregions (of 4×4 pixels). An orientation histogram of 8 bins is created from each subregion. These histograms are computed from magnitude and orientation values of the pixels in these subregions. The gradient orientations are reported to the main orientation computed before, to provide the invariance of the descriptor to rotations. The magnitudes are further weighted by a Gaussian function (see figure 5.10). The final SIFT descriptor is a 128 dimensional vector obtained by

concatenating all these histograms ($4 \times 4 \times 8 = 128$). This vector is then normalized to unit length in order to enhance invariance to affine changes in illumination. To reduce the effects of non-linear illumination a threshold of 0.2 is applied and the vector is again normalized.

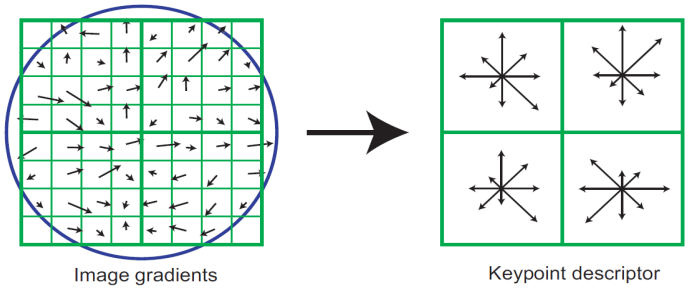


Figure 5.10: A keypoint descriptor is created by first computing the gradient magnitude and orientation at each image sample point in a region around the keypoint location, as shown on the left. These are weighted by a Gaussian window, indicated by the overlaid circle. These samples are then accumulated into orientation histograms summarizing the contents over 4x4 subregions, as shown on the right, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region. This figure shows a 2x2 descriptor array computed from an 8x8 set of samples, whereas the optimal and described optimal descriptor use 4x4 descriptors computed from a 16x16 sample array (source [Lowe 2004]).

SIFT descriptor is highly distinctive and partially invariant to the remaining variations such as illumination and 3D viewpoint (affine transformation).

5.2.2 SIFT Feature Detection and Selection For Object tracking

The detailed SIFT descriptor can describe any 16×16 square region of image, independently of the existence of a point of interest on its center or not, but computing it around a point of interest (SIFT point) is more efficient for matching purpose.

We want to exploit the effectiveness and the discriminative power of SIFT descriptors in our object tracking algorithm, but due to the related constraints to the addressed context (video surveillance) like small objects, low resolution and noisy images, etc. the SIFT point detection method described below does not always ensure a sufficient number of SIFT points to describe the whole object of interest, and their localisation is not necessary uniformly distributed (see figure 5.11 (b)). This detection can also provide too many points, and tracking all these points increase the required processing time.

For these reasons, we use a modified method for SIFT point selection for the object representation. First, a more permissive SIFT point detection is performed, by avoiding

low contrast and edge response filtering. More SIFT points are then detected (see figure 5.11 (c)).

Figure 5.11: SIFT points detection results. (a) The original low resolution image. (b) Detected SIFT points using the whole detection process. 37 points are detected on the image, 16 of them are on the object of interest (the person). (c) Detected SIFT points without low contrast and edge response filtering. More points are detected: 74 points are detected on the image, 39 of them are on the object of interest (the person).

Once a large number of points of interest is detected inside the bounding box which delimits the object of interest, only foreground points of interest are kept. This is performed using the object mask provided by the background subtraction algorithm (see figure 5.11 (b)).

Finally, object image is divided into subregions, in which a constant number n of SIFT features is kept and the other points are rejected (see figure 5.12 (c)). If a given subregion contains more than n points, the most reliable of them, in terms of contrast and curvature values are selected. In fact, even if these two values are not used during SIFT point detection for candidate point filtering, we use them to sort detected points in each subregion and to select the most reliable of them.

The object image subdivision can be performed in several subregion sizes. We have tried many configurations and have seen that for people, a grid of 4×6 subregions provides the best compromise between tracking performance and processing time. The same remark can be done for the number of selected SIFT points in each subregion. When the subregions are relatively small (like the considered 4×6 subregion subdivision for video surveillance images), using one point per subregion provides slightly less good results than using two points per subregions, but the processing time is multiplied by 2 for a non significative improvement in performances, especially after performing our data association framework, which will be detailed in next sections, and which compensates for the slightly decreased performances provided by only one point per subregion.

Finally, the object is represented with a sufficient number of SIFT points, uniformly

distributed on the image (see figure 5.12 (d)), allowing better partial occlusions management as we will explain it in next sections. For people images, the number of representative points generally varies from 15 to 20 points.

Note that even a subset of selected points are less reliable due to the non-filtering of low contrast and edge points, they are still better than randomly selected points in SIFT points-free regions. The aim is to describe uniformly distributed local regions on the person. The low reliability level of these points is compensated by the data association method described few sections below.

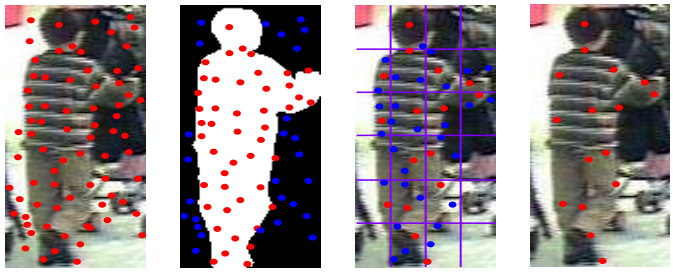


Figure 5.12: SIFT point selection for object representation. Red points are the ones selected and the blue ones are the ones removed at each step. (a) All the detected SIFT points without low contrast and edge response filtering. (b) Background point of interest removed by background subtraction mask. (c) Object of interest subdivision and selection of constant number points of interest per subregion, by keeping the most reliable ones. (d) Final object of interest (person) representation using SIFT points

5.3 SIFT Feature Tracking By Particle Filtering

After object of interest detection and modeling using SIFT features, the first level of tracking is performed on the SIFT features only. At this level, the tracking algorithm does not care about to which object belongs a given SIFT point. It tracks all SIFT features as independent entities. These points are tracked over time using a specific particle filter.

As a reminder about Bayesian filters, detailed in sec. 2.2.2.2, let x_t denote the state of the system at the current time t, and $y^t = (y_1, ..., y_t)$ the observations up to time t. The filtering problem involves the estimation of the state vector at time t, given all the measurements (observations) up to and including time t. In a Bayesian setting, this problem can be formalized as the computation of the distribution $p(x_t|y_{1:t})$, which can be done recursively in two steps.

prediction

In the prediction step, $p(x_t|y_{1:t-1})$ is computed from the filtering distribution $p(x_{t-1}|y_{1:t-1})$

at time $t-1$:

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1} \tag{5.7}$$

where $p(x_{t-1}|y_{1:t-1})$ is assumed to be known due to recursion. The distribution $p(x_t|y_{1:t-1})$ can be thought of as a prior over x_t before receiving the most recent measurement y_t.

update

The previous prior is updated with the new measurement y_t using the Bayes' rule to obtain the posterior over x_t:

$$p(x_t|y_{1:t}) \propto p(y_t|x_t)p(x_t|y_{1:t-1}) \tag{5.8}$$

In general, the computations in the prediction and update steps (eq. 5.7 and 5.8) cannot be carried out analytically, hence the need for approximate methods such as Monte Carlo sampling, also called Particle Filtering methods. Here, the space of hypothesis is explored using a set of particles, which are projected in the prediction step, and sampled by their importance in the update step, allowing to estimate the new system state.

The recursion requires the specification of a dynamic model describing the state evolution $p(x_t|x_{t-1})$, and a model giving the likelihood of any state in the light of the current observation $p(y_t|x_t)$. The recursion is initialized with some initial distribution $p(x_0)$.

In our approach, the state of a SIFT feature $\mathbf{x} = \{x, y, u, v, \mathbf{h}, n\}$ consists of:

- The SIFT feature position (x, y).
- The velocity component (u, v).
- The SIFT descriptor $\mathbf{h}$ associated to the SIFT point.
- The measurement error estimation n following a normalized distribution.

We have tested the use of acceleration component as an additional state information, but no significant improvement has been observed. This is due to the constant or slow variations of objects velocity.

In particle filtering, each hypothesis about the new state is represented by a particle which has its own state with the same structure as the SIFT feature one. Each SIFT feature is then tracked using a constant number of particles.

The prediction step consists in applying the dynamic model of each SIFT feature to all its associated particles to compute the new estimated location of each particle (see figure 5.13). This is done for each particle by:

$$(x,y)_t = (x,y)_{t-1} + (u,v)_{t-1}.\Delta_t + n_{t-1}(x,y) \tag{5.9}$$

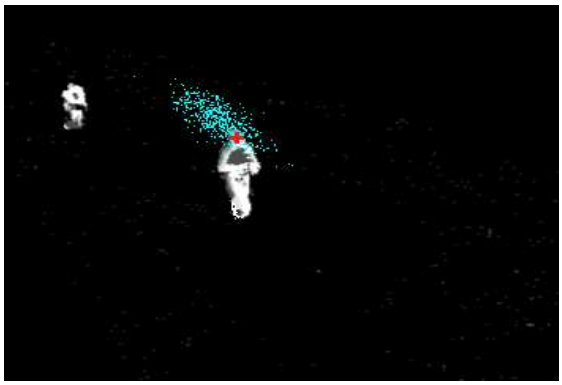


Figure 5.13: Prediction step: The red point is the previous location of the SIFT point to track at time $t-1$, and the cyan points represents the predicted positions of all the particles at the current time t

The update step consists in estimating the new location of the tracked feature using the predicted state of all particles and some measurements provided by the current image at time t. This step is performed in three sub-steps: particle weighting, particle sampling-resampling and new state estimation.

5.3.1 Hybrid Particles Weighting

Generally, when a given descriptor is tracked using a particle filter, the particle weighting is performed using a similarity measure between the tracked descriptor and each particle descriptor.

In our case, we have followed this method in a first time. For a given tracked SIFT point, a SIFT descriptor is computed around each associated particle. The weight of a given particle is then provided by:

$$W_p = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d(H_f,Hp)^2}{2\sigma^2}} \tag{5.10}$$

where

$d(H_f, H_p)$ denotes the similarity between the SIFT descriptor of the tracked point (H_f) and SIFT descriptor the considered particle descriptor (H_p). We use an Euclidean distance after having tested some other distances without getting significant improvements.

σ denotes a standard deviation computed on the tracked feature similarity variations up to time $t-1$.

The tracking of SIFT points using this weighting method for particle filtering performs well in most of cases, but we have observed a critical issue for some tracked SIFT points in some situations. Due to the small size of object images provided in video surveillance context, and to the low resolution of images, when a SIFT point is located too close to the object contours, and if the described region and the nearest background are low textured, the tracked SIFT can leave the object of interest and cling on the background because most of particles are located on the background after prediction (see figure 5.14).

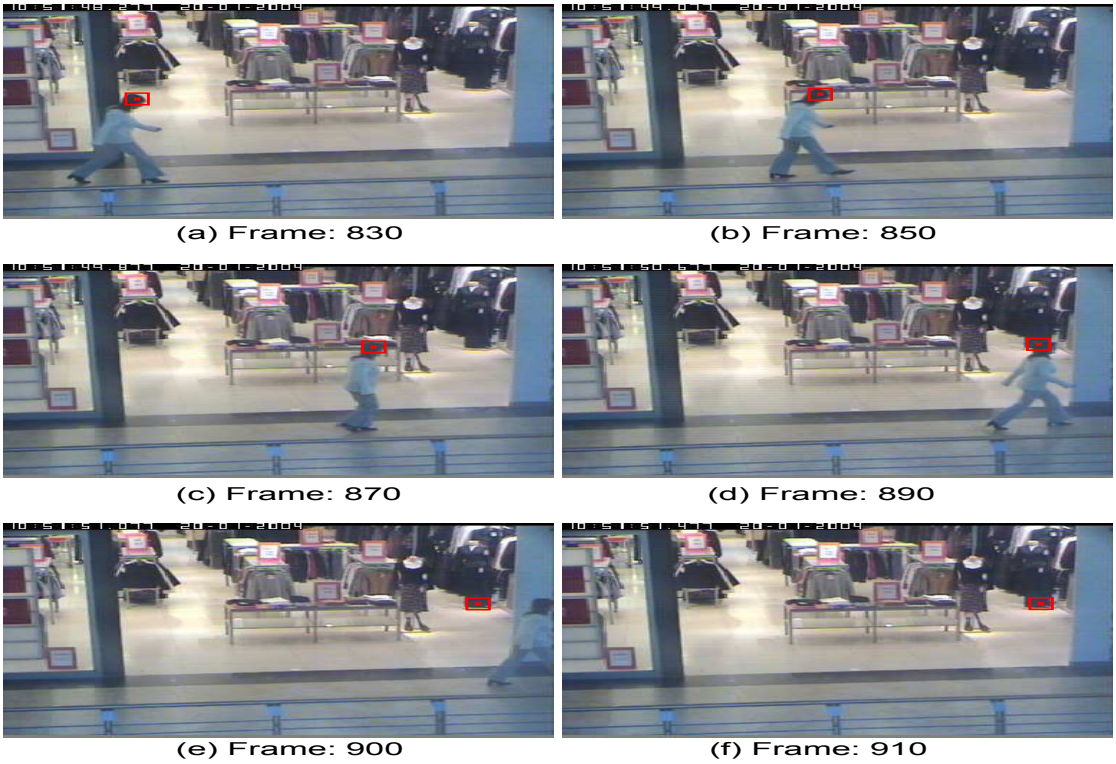


Figure 5.14: SIFT point tracking failure due to small size of the person's image, the proximity of the SIFT point to the person contour and to the low texture on the containing region. The SIFT point (red cross) is tracked correctly until frame 890, where it clings on the background which presents similar descriptor region (red square) at this location

We deduce that the particle weighting by only similarity measure between SIFT descriptor of the tracked point and those of the particles is not sufficient. It is necessary to deal with background proximity during the weighting, especially when particles are scattered to cover a maximum amount of likely hypotheses and when a large amount of

them are on the background (see figure 5.13).

The intuitive idea can be to use the background subtraction results to directly avoid particles which are on the background by assigning a null weight to them. This idea is risky because the quality and the reliability of background subtraction algorithm may greatly vary according to the scene context (contrast, illumination, etc.) and to the background subtraction algorithm itself (the used method, parameters, etc.). Thereby, using a binary weighting, i.e. 1 for foreground particle and 0 for background one, is not a satisfactory solution. We have observed that using this binary weighting causes incorrect SIFT point tracking according to the background subtraction quality. When the extracted foreground pixels are too low, like in figure 5.15(b) (low contrast and hight thresholds in the background subtraction algorithm), the new localisation of the SIFT point in each new frame slides on the foreground pixels and varies greatly. The localisation is then not precise.

We propose a method to take into account background subtraction results to assign a real foreground weight to particles, even if the background subtraction provides a binary separation between background and foreground. This method allows to deal with the diverse background subtraction qualities. By modifying some parameters in the used background subtraction algorithm (contrast thresholds and the number of considered gaussian per pixel), we have obtained three different background subtraction qualities (see figure 5.15).

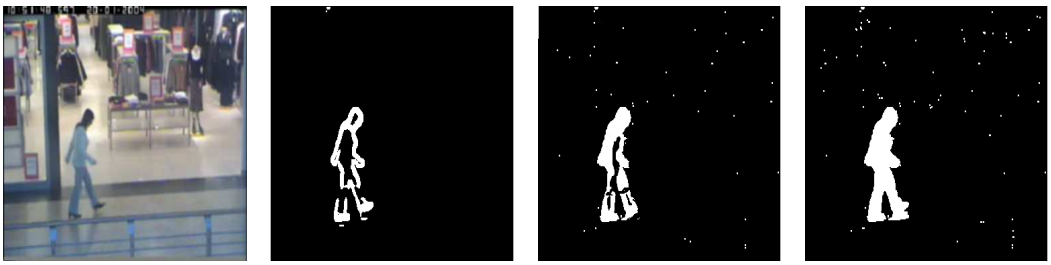


Figure 5.15: Different qualities of motion detection. (a) Original image. (b) Low detection. (c) Medium detection. (d) High detection

The particle weighting used in our tracking algorithm is then given by:

$$W_p = \frac{\mathbf{c}}{\sigma\sqrt{2\pi}} e^{-\frac{d(H_f,H_p)^2}{2\sigma^2}} \tag{5.11}$$

where $\mathbf{c} \in \{c_0, 1\}$ is the foreground probability weight.

If the corresponding pixel to the considered particle is an indication of a foreground one by the background subtraction result, $\mathbf{c} = 1$. Otherwise, $\mathbf{c} = c_0$.

c_0 is computed for each detected object at each frame as:

$$c_0 = \cos(\rho\frac{\pi}{2}) \tag{5.12}$$

and:

$$\rho = min\left\{1, \frac{density}{density_0}\right\} \tag{5.13}$$

where **density** is the ratio between the number of foreground pixels inside the object bounding box and the area of this bounding box, and $\mathbf{density_0}$ is the approximate real density of the foreground pixels of an object inside its bounding box. For a given type of objects, **density** is a variable value, provided at each frame by the background subtraction results while $\mathbf{density_0}$ is constant and estimated off-line using some ground truth on the considered object type. For example, for a walking person, the average $density_0$ is approximately 0.5 (it means that in the minimal bounding box of a walking person image, approximately 50% of the pixels inside the bounding box belongs to the person).

Let us consider the possible situations of several background subtraction qualities at both localisation of a given particle: in the foreground and in the background. In all these cases, if the particle is on a foreground pixel, $c = 1.0$. Here, the similarity measure between the SIFT descriptor of the tracked SIFT point and the one of the particle will decid the association. The ambiguity occurs for background pixels.

- For a low and medium foreground pixel detection (figure 5.16 (a,b)), the value of ρ increases with the increase of the foreground extraction quality, and thereby, particles on the background obtain decreased weight c_0 but are not discarded, to avoid the loss of useful information.

- For a good foreground pixel detection (figure 5.16 (c)), the value of ρ is near 1.0 and thereby, the value of c_0 is near 0. It means that the more the background subtraction is precise, the less the particles on background are important (lower weight).

- For an excessive foreground pixels detection (figure 5.16 (d)), the value of ρ is equal to 1.0 and thereby, the value of c_0 is near 0. It means that an estimated background pixel in this case has a high probability to be a real background pixel. Real background pixels which are detected as foreground ones in this case will have a value of $c = 1.0$, but does not alter the particle weighting as long as this weighting is not worse than a standard weighting method which is only based on similarity measure (eq. 5.10), but equal to it.

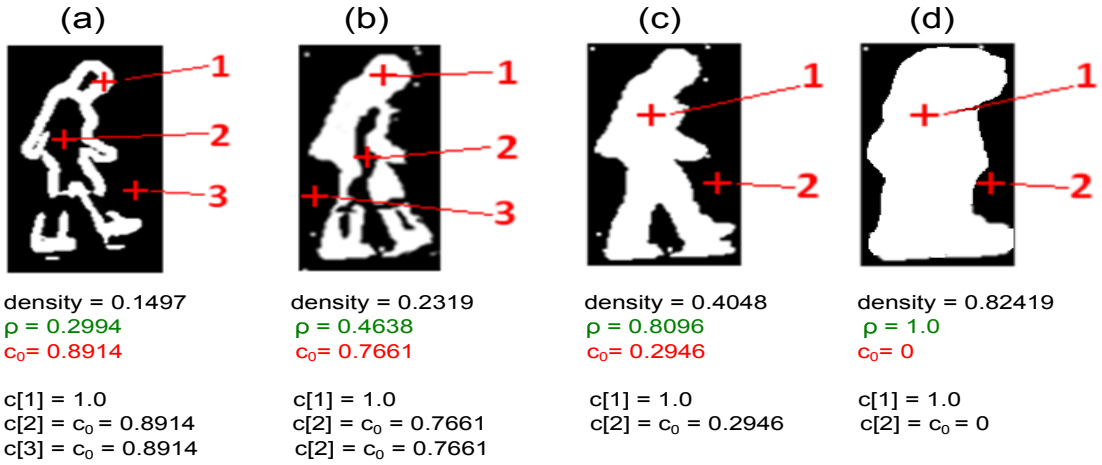


Figure 5.16: Hybrid particle weighting.
Hybrid particle weighting, taking into account background subtraction result qualities. The $density_0$ value is taken as $density_0 = 0.50$. (a-c) From low to good foreground extraction, by varying the contrast thresholds in the background subtraction algorithm. (d) Excessive (erroneous) foreground pixel extraction by decreasing contrast threshold and considering the mean of large blocks of pixels during the background subtraction

By using this new weighting method, previously lost tracked points, due to the explained phenomenon, are tracked correctly.

Note that we have chosen to take a cos function to encode the transition from the lower background subtraction quality to higher one smoothly (eq. 5.12). We have tested a linear function to perform this task and we have obtained satisfactory results, but the transition is encoded roughly and the precision is lower than the cos function one.

This test consists in annotating manually the localisation of a given point on the head of the walking person on each frame of a small video sequence, providing a ground truth of the real localisation of the same point in each frame. We have tested both linear and cos function by automatically tracking the same point along this sequence, by varying the background subtraction parameters, and by computing localisation error during each

tracking. This error is taken as the mean of distances between the estimated position of the point by the tracking and its real position (ground truth one). We have observed that this error is lower with the cos function (see table 5.1).

	cos	linear
Low detection quality	5.7	6.1
Medium detection quality	5.2	5.3
High detection quality	3.4	3.6

Table 5.1: Comparison between the cos function and a linear function for background quality transition encoding. The mean error of point position estimation (in pixels) is lower for cos function.

5.3.2 Particles Sampling and Resampling

After weighting, all particles are sampled using a "Sampling Importance Re-sampling" (SIR) method [Tanner 1987, Smith 1992] to keep the most important ones, drop the less important and replace them by new particles generated from the kept ones. The sampling step allows the tracker to keep the more reliable particles and the re-sampling step avoids information degeneration, as explained in sec. 2.2.2.2 (Particle filters paragraph). Each feature keeps a constant number of particles over time, which makes the processing time easier to control. Finally, all particles are re-weighted with the same normalized weight.

5.3.3 New State Estimation

The estimation of the new location of the tracked feature is obtained as the centroid of all its particles (see figure 5.17). The descriptor of the tracked feature is computed around the new location.

A variation measure between the previous descriptor and the new one is computed for each tracked point. This variation measure is used for the feature variation learning in a Gaussian model in order to decide if a new state is acceptable or not (if it fits the Gaussian model variations according to the standard deviation). If the variation is too important the SIFT point is discarded and replaced by a new detected point in the same subregion as the discarded one (see 5.2.2 for subregion division). Otherwise, the point is kept and a reliability measure γ is computed using the variation measure as:

$$\gamma_t = 1 - \frac{\text{variation measure at time t}}{\text{max accepted variation measure in the Gaussian model}} \tag{5.14}$$

This reliability measure γ is used for link weighting during the data association step, as described in the next section.

Figure 5.17: Tracking of one SIFT point on the head of a person by particle filter: (a) the predicted position of all particles in blue and yellow. The yellow particles are the sampled ones, which will be used to generate new ones to replace degenerated particles in blue. (b) in red, the new location of the tracked SIFT point, computed as the barycentre of all weighted sampled particles.

After the update step, the velocity $(u, v)_t$ and the measurement error estimation $n_t(x, y)$ components of each SIFT feature are also updated. For each SIFT feature, a linear regression function is computed on the p last localisations of the considered SIFT features. The regression line direction provides the estimated direction of motion for next time $t+1$ (velocity vector direction), the mean of displacement magnitude between these successive p positions provides the motion velocity magnitude, and the variance on these displacements provides the measurement estimation error for the next time $t+1$. In our experiments, the optimal value for p is 10 (see figure 5.18(c)). We have observed that for lower values, the regression line direction, the mean and the variance of displacements vary too fast due to "non-smooth" SIFT point trajectories (see figure 5.18(a)). In the other hand, for large values of p, the regression line direction takes more time to follow the changes in motion directions (see figure 5.18(b)) causing tracking failure du to poor prediction process.

Note that the value $p = 10$ is extracted experimentally, and is not the optimal value for all situations. The optimal value depends strongly on the movement, but statistically, a medium value between 7 and 10 provides the best tracking results in most of the cases.

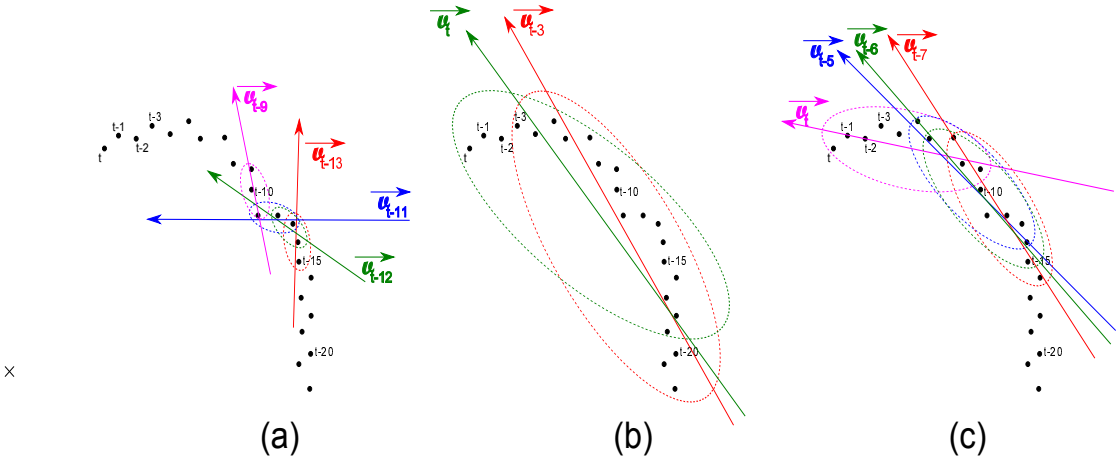


Figure 5.18: Velocity component update using regression function performed on various set of last estimated positions. The 22 last positions of a tracked SIFT point are used for illustration. (a) Only the 3 last positions are used for regression computation. The resulting velocity vector is too sensitive to local variations and the prediction at next time are erroneous. (b) The 20 last positions are used for regression computing. The resulting velocity vector does not follow movement direction variations quickly. at time t, the estimated movement direction (green vector) indicated that the movement direction is to the top-left side while the real movement direction is to the left (see the 8 last positions). (c) The 10 last positions are used for regression computing. In this case, movement direction changes are managed better than in the two previous cases

5.4 Data Association

At the end of previously detailed step, all the SIFT points of all the considered objects at time "*t-1*" are localised in the current video frame. Data association step consists in linking previously tracked objects at time "*t-1*", noted "*to(t-1)*", with the new detected objects at the current frame "*t*" , noted "*do(t)*", while dealing with complex situations like partial or full occlusions.

From the previous frame (at time "*t-1*") to the current one (at time "*t*"), only five general cases can occur (see figure 5.19):

- In the first case, called "1 to 1" correspondence (figure 5.19 (a)), a unique detected object "*do(t)*" corresponds to only one previously tracked object "*to(t-1)*". This is the simplest case. Here the algorithm updates the "*to(t-1)*" localisation at time "*t*" by linking it directly to "*do(t)*".

- In the second case, called "N to 1" correspondence (figure 5.19 (b)), a unique detected object *do(t)* corresponds to a set of Q tracked objects $\{\mathrm{to_k(t-1)}\}_{\mathrm{k:1...Q}}$. This situation occurs when the detection at time "*t*" did not correctly split detected

moving objects, typically during partial occlusions or high object proximity. Here the algorithm splits the bounding box of "*do(t)*" into Q smaller bounding boxes using the spatial distribution of the SIFT points before the merge. This distribution is given by the ratios between feature locations and the borders of the bounding box when the objects were separated.

- In the third case, called "1 to N" correspondence (figure 5.19 (c)), a set of *R* detected objects $\{do_l(t)\}_{l:1...R}$ corresponds to a unique tracked object "*to(t-1)*". This situation occurs during the dispersion of a group of objects or at the end of occlusion between objects of interest. Here two situations can be distinguished:
  - "$to_i(t-1)$" can be the result of a previous merge (occlusion) of tracked objects at time "*t-p*" as described in the previous case (b). In this case, the tracking is resumed using the occlusion management approach described in next section (5.4.2).
  - "$to_i(t-1)$" has always been tracked as a group of objects since its appearance in the scene, new tracked objects $\{to_l(t)\}_{l:1...R}$ are initialized by each $\{do_l(t)\}_{l:1...R}$ after the split.
- In the fourth case, called "1 to 0" correspondence (figure 5.19 (d)), no detected object "*do(t)*" corresponds to the tracked object "*to(t-1)*". This occurs in full occlusion situations or when the tracked object "*to(t-1)*" leaves the scene at time "*t*". The two cases are managed differently:
  - If the tracked object is close enough to a scene exit (image borders or known exit zones in the image) at time "*t-1*", the tracking algorithm considers that this object has left the scene and stops its tracking definitely.
  - In the other case, the tracked object "*to(t-1)*" is considered as in occlusion situation. Here, the tracking algorithm stores the lost object for tracking recovery if it re-appears later. This process if performed using the occlusion management approach described in next section (5.4.2).
- In the last case, called "0 to 1" correspondence (figure 5.19 (e)), a detected object "*do(t)*" does not correspond to any previously tracked object "$to_i(t-1)$". This situation occures either when a new object apprears in the scene for the first time or when an previously tracked object has been occluded few time before and reappears. In the case of new object appearence, a new tracked object is initialisez. In the case of occluded object reappearance, the occlusion managment resume its tracking as described in the occlusion management section (sec. 5.4.2).

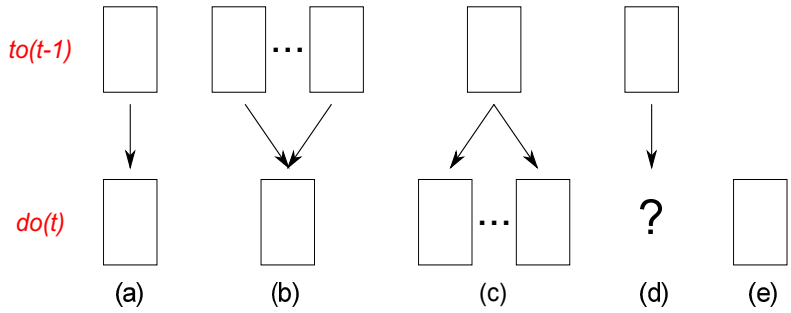


Figure 5.19: The five possible cases during the object tracking from a previous frame to the current one. (a) A simple 1 to 1 correspondence. (b) Merged tracked objects due to high object proximity or partial occlusion. (c) Split tracked object due to grouped object separation or occlusion end. (d) Lost tracked object due to scene exit or full occlusion situation. (e) New detected object without correspondance with any tracked object in previous frame. It can be a new appearing object or a priviously occluded one.

5.4.1 Case Identification

The first step of our data association method consists in identifying in which case each tracked object "*to(t-1)*" is at time "*t*" before managing the case.

To do this, an $M \times N$ link score matrix, denoted **S**, is constructed. M is the number of tracked objects "*to(t-1)*" and N is the number of detected objects "*do(t)*".

Each element $s(i,j)$ of **S** is calculated as the weighted proportion of SIFT points from the i^{th} tracked object "*to(t-1)*" that geometrically belongs to the j^{th} detected object "*do(t)*" using the following formula:

$$s(i,j) = \frac{1}{\sum_{q=1}^{p} \gamma_q} \sum_{k=1}^{p} \gamma_k(i,j) \tag{5.15}$$

where p is the number of SIFT points of the tracked object "$to_i(t-1)$" which belongs geometrically to the detected object "$do_j(t)$", $\gamma_k(i,j) \in [0,1]$ is the reliability measure of the k_{th} tracked SIFT point, computed at the end of particle filtering update step (eq. 5.14). The contribution of each SIFT point in the link score value is directly proportional to its reliability, providing a better temporal linking process.

Putting these link score values in a matrix form eases and speeds up the decision process. We use the Hungarian algorithm [Kuhn 1955] to select the best links, i.e. we do not take the absolute best links one by one, but we select the links which provides the best global score. The case identification is performed using the link scores.

- "1 to 1" correspondence is represented by a high link score ($\approx$ 1.0) between a tracked object and a detected one.

- "N to 1" correspondence is represented by a set of N close link scores ($\approx 1/N$) between a N tracked object and one detected one.

- "1 to N" correspondence is represented by a set of N close link scores ($\approx 1/N$) between a tracked object and a set of N detected one.

- "1 to 0" correspondence is represented by low link scores (≈ 0.0) between a tracked object and all the detected ones.

- "0 to 1" correspondence is represented by remaining columns in **S** after linking, i.e. detected objects "$do_j(t)$" which are not linked with any tracked objects "$to_i(t-1)$".

Note that after this data association step, SIFT points outside of their objects (moved onto the background or onto other objects during occlusions) are dropped and replaced by new detected SIFT features following the detection and selection process described in sec (5.2.2). Sub-regions which are common to multiple objects in the case of partial occlusions are not used for the detection to avoid ambiguous situations.

On the other hand, the tracking algorithm keeps a uniform spatial repartition of the SIFT features by filtering out too close features. The system keeps the most reliable feature and replaces others by new detected ones in sub-regions of the object with no/fewer SIFT points.

Figure 5.20: From SIFT point tracking to Object (Person) tracking: (a) four SIFT points at different location of the person body are tracked using particle filter. (b) The person tracking result after data association step.

5.4.2 Occlusion Management

The partially occlusions are handled by a continuous tracking process on the remaining visible SIFT points. If the partially occluded object became fully visible, new additional SIFT points are detected and assigned to the tracked object using the same process as the one described for SIFT points detection and selection. This is due to maintain a global and well distributed representation of the object.

After link creation using **S**, some detected objects "*do(t)*" may not be linked with any "*to(t-1)*". They can corresponds to new objects appearing for the first time in the scene or previously occluded objects which re-appear.

Before initializing new tracked objects "*to(t)*" with unlinked "*do(t)*", an attempt to match these unlinked detected objects "*do(t)*" with tracked objects in occlusion state is performed using two reacquisition methods. The first one is fast and requires some basic information, extracted during the tracking of objects. The second one is more sophisticated and uses the visual signatures of the occluded objects to re-identify them. This visual signature computing is described in the re-identification chapter (chapter 6). The second reacquisition method is performed only if the first one provides ambiguous result, i.e. if the matching is not validated and not rejected with reliable decision (score). Most of the time, the first method is sufficient to manage the occluded object tracking recovery. It is performed as follow:

During the tracking process of a fully visible object, four kinds of information are extracted :

- The last state of its associated SIFT points (the point localisations on the object and their descriptors) are stored from the last fully visible image of the object.
- The variations of its estimated real world dimensions (W and H), extracted using the camera calibration matrix as explained in (sec 5.1) are encoded in a Gaussian model.
- The variations of its estimated real world velocity on the ground plane, also extracted using the camera calibration matrix, are encoded in another Gaussian model.
- Its n dominant colors in HSV space are stored. We take only Hue value in consideration. n depends on the class of tracked object. For people, we use $n = 2$ (due to the general separation of a person body in torso and legs) while we use $n = 1$ for vehicles.

A first matching attempt is performed using SIFT descriptors matching. A set of SIFT points are detected and their SIFT descriptors computed on the detected object. A

standard point to point matching is performed. If the SIFT point matching is successful, the reacquisition is validated at this point and no other matching process is required. This is due to the high discriminative power of SIFT descriptors. On the other side, if SIFT point matching fails, the object matching process continues. This is due to the point of view dependence of detected SIFT points. The occluded object may reappear with a different visible side than the one with which it disappears and on which the stored SIFT points have been detected.

A second matching rejection process is performed using real object dimensions. If the dimensions of the detected object "*do(t)*" does not fit in the computed Gaussian model of the occluded object (using the variance of dimensions of the Gaussian model), the matching is rejected.

Finally, a matching score between a detected object "*do(t)*" and an occluded one "*to(t-p)*" is computed by:

$$score = \frac{1}{3}(Score_W + Score_H + Score_{hue}) \tag{5.16}$$

where:

$$Score_W = 1 - \frac{(\text{detected object W} - \text{occluded object mean W})^2}{\text{max observed W variation}} \tag{5.17}$$

$$Score_H = 1 - \frac{(\text{detected object H} - \text{occluded object mean H})^2}{\text{max observed H variation}} \tag{5.18}$$

$$Score_{hue} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|\theta_1^{(i)} - \theta_2^{(i)}|}{180}\right) \tag{5.19}$$

n is the number of considered dominant colors, θ_1 is the corresponding angle in Hue color disc to the dominant hue value of detected object, and θ_2 is the corresponding angle of the dominant hue value of occluded object. These angle values are taken in degrees.

Two remarks are to be done for this matching score.

The first one concerns the uniform weighting of each information. In future work, an importance weighting method may be proposed according to the tracking result observations.

The second remark concerns the non use of velocity information in this matching score. In fact, a detected object "*do(t)*" does not have a velocity as it is a static object at the current frame and no temporal information are available for it. This velocity information will be used later.

High matching scores are used to validate occluded object reacquisition.

Non linked detected objects in the current frames, either by continuous tracking or by the first reacquisition attempt, are used to initialize new object tracking since they are considered as new appearing objects in the scene.

After few frames of tracking these new object, a second attempt for occluded objects reacquisition is performed using the velocity information. If the computed velocity of the new tracked object during few frames does not fit the velocity Gaussian model of occluded object, the matching is avoided. Otherwise, a new matching score is computed by:

$$score = \frac{1}{4}(Score_W + Score_H + Score_{hue} + Score_{velocity}) \tag{5.20}$$

where:

$$Score_{velocity} = 1 - \frac{|\text{new tracked object velocity} - \text{occluded object velocity}|}{\text{max observed velocity variation}} \tag{5.21}$$

Note that each occluded object is stored for requisition purpose only for a predefined time, to avoid the combinatorial explosion which can occur when the number of occluded object to reacquire increases greatly.

In our experiments, we start this second matching attempt, based on the new object velocity, after 50 frames of tracking to have a stable and reliable object velocity to compare, and we keep occluded objects for a maximum time of 1 mn before definitely considering them as lost (out of the scene).

This method allows to deal with most of occlusion situations, and is validated by the experimental results provided in chapter 7

5.5 Conclusion

We have proposed an object tracking algorithm, based on SIFT features for object representation, particle filtering for SIFT point tracking, and a data association framework to achieve the object tracking reliably.

The use of SIFT features is justified by two main reasons: the sparse point representation of the object allows more flexible tracking and object deformations/partial occlusion management. The robustness of SIFT descriptors and their high discriminative power increase the reliability of the tracking.

Particle filtering for SIFT point tracking draws its interest in the ability of paralleled exploration of several hypotheses for the new localisation of the tracked SIFT point. The contribution which we have proposed for particle weighting and SIFT point reliability measures allows a more reliable tracking.

The final data association framework allows to detect all possible kinds of situation which can occur during object tracking, especially the occlusion case. A Weighted temporal linking process is proposed, achieving the visible object tracking.

For occluded objects, we have proposed a real time method to perform their reacquisition if they reappear in the scene.

A modified version of our tracking algorithm, using FAST interest points and HOG descriptors around them, has been integrated in the main intelligent video surveillance software of Digital Barriers company, called "SafeZone®", which has been deployed on many video surveillance systems in the world since 2011. It provides better tracking performance in comparison with the previous used tracking algorithm. This validate also the genericity of the proposed algorithm to any kind of local descriptor computed arround some points of interest.

The evaluation and comparison with state of the art results which validate our tracking approach is be provided in chapter 7, but some issues still exists with our tracking method. The first one is the use of image intensity information only during SIFT point tracking. The use of color SIFT descriptor will be tested in future work. The use of other type of information (local/global color descriptors, covariance descriptors, etc.) is also considered and will be tested in future work.

6

FAST PEOPLE RE-IDENTIFICATION

Color and texture are two important pieces of information for people appearance modelling. We have built our re-identification technique based on Farenzena et al. approach [Farenzena 2010]. This approach can be used both for single-shot and multi-shot cases, the . We present both cases but we focus more on multi-shot one due to the availability of mono-camera tracking results and the better results of re-identification when multiple images of each person are used.

Our choice to take [Farenzena 2010] as work basis is motivated by the possibility to apply it in both single and multiple shot cases and by the fast processing (real time processing for small sampled set of images per person). The original approach ([Farenzena 2010]) provides interesting performances but has some issues. Some other recent approaches provide better re-identification results like MRCG ([Bak 2011]), CPS ([Cheng 2011]) and LMNN-R ([Dikmen 2011]) (see chapter 7) but are more constrained or are inadequate for real deployed video-surveillance systems (as we will see in evaluation chapter 7, some popular datasets for re-identification evaluation does not really match real video-surveillance requirements and constraints). For example, MRCG ([Bak 2011]), is highly time consuming (it requires 6 s in average to compute a visual signature of a person using 46 images, more details are available in chapter 7) due to the covariance means computation (eigenvalues decomposition and gradient descent iterations) computed on many grid cells (after dividing human body into a grid), and it provide a hardly updatable signature, i.e. the computed signature of a person with "n" images cannot be updated with a new image "n+1" (because of gradient descent iteration for mean covariance computing), and requires to recomputed the signature with the "n+1" images as a new signature if we want to use new acquired images during

processing.

We start by presenting the standard approach proposed in [Farenzena 2010]. After that, we discuss the main issues of this approach, some of them are also issues for more recent and efficient approaches. Finally, we detail our contribution to solve these issues, achieving an improved version of [Farenzena 2010] approach which provides comparable/slightly better results than state of the art ones while keeping the real-time processing of the original approach and dealing with more complex situations (object rotations, etc.).

6.1 Person Re-identification by Symmetry-Driven Accumulation of Local Features

The basic approach requires a background/foreground separation obtained by a background subtraction algorithm in the multi-shot case, or by using STEL (Structure Element) technique [Jojic 2009] in case of single-shot. It computes a visual signature which is exclusively based on color information. This signature consists of three different representations of the color information, each representation focuses on specific aspect of the frequency/localisation of the color. To make the representation more robust to rotations and partial occlusions, the image of a human body is divided into 4 parts, according to two symmetry and two asymmetry axes. The extracted color descriptors are then reported to the body part to which they belong

6.1.1 Body subdivision: Assymmetry and Symmetry Axes

To perform the body image subdivision, the two horizontal asymmetry axes which separate head from torso, and torso from legs are firstly searched. Once they are defined, two other vertical symmetry axes, one dividing the torso and the other dividing legs in two symmetric parts are searched.

Farenzena et al. [Faranzena] define two operators to perform this task. Given a person image of width J and height I, the first operator is the chromatic bilateral operator defined by:

$$C(i,\delta) = \sum_{B[i-\delta,i+\delta]} d^2(p_i, \hat{p}_i) \tag{6.1}$$

where $d(.,.)$ is the Euclidean distance, evaluated between HSV pixel values p_i,$\hat{p}_i$, located symmetrically with respect to the horizontal axis at height i. This distance is summed inside a vertically sliding rectangular window $B[i-\delta, i+\delta]$, with the same width

J, and a height of 2δ centred vertically on the height i. To achieve the independence to scale changes, δ is taken as $I/8$ (See figure 6.1).

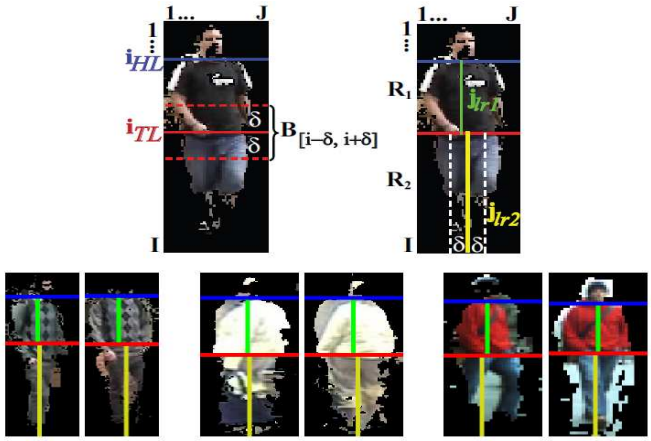


Figure 6.1: Symmetry-based Silhouette Partition. On the top row, overview of the method: first the asymmetrical axis i_{TL} is extracted, then i_{HT} ; afterwards, for each Rk region the symmetrical axis j_{LRk} are computed. On the bottom row, examples of symmetry-based partitions on images from the datasets. As you can notice, they coherently follow the pose variation. (source [Farenzena 2010]).

The second operator is the spatial covering operator, which calculates the difference of foreground areas for two regions, defined by:

$$S(i,\delta) = \frac{1}{J\delta}|A(B[i-\delta, i]) - A(B[i, i+\delta])| \tag{6.2}$$

where $A(B[i-\delta, i])$ is the foreground area in the window of width J and height δ, delimited by $[i-\delta, i]$.

Using these two operators (eq. 6.1 and eq. 6.2), the four axes (2 asymmetry and 2 symmetry axes) can be estimated as follow:

- The asymmetry axis i_{TL} which separate torso from legs is defined as:

$$i_{TL} = \underset{i}{\operatorname{argmin}}(1 - C(i,\delta) + S(i,\delta)) \tag{6.3}$$

  this provides the horizontal axis which separates two neighbouring regions with strongly different appearance (colors) and similar areas. The values of C are normalized. To reduce searching time and avoid some possible errors, the search of i_{TL} holds in the interval $[\delta, 1-\delta]$.

- The asymmetry axis i_{HT}, separating the head from the torso is defined as:

$$i_{TL} = \underset{i}{\operatorname{argmin}}(-S(i,\delta)) \tag{6.4}$$

  this provides the horizontal axis which separates two neighbouring regions which strongly differ in area. For the same reason as previously, the search for i_{HT} is performed in the interval $[\delta, i_{TL} - \delta]$.

- The two symmetry axes separating torso and legs horizontally are defined by:

$$j_{LRk} = \underset{i}{\operatorname{argmin}}(C(j,\delta) + S(j,\delta)) \tag{6.5}$$

  where $k = 1,2$ corresponds to the regions R1 and R2, respectively the regions defining the torso and the legs. Here, the searching of the torso symmetry axis is performed in the region delimited vertically by $[i_{HT}, i_{TL}]$, and the legs symmetry axis in performed in the region delimited vertically by $[i_{TL}, I]$, both by a sliding window in interval $[\delta, 3\delta]$, $\delta = J/4$ for this time.

Note that the head part (delimited vertically by $[0, i_{HT}]$ is ignored in this approach due to the poverty of extractable information from this region in video surveillance context (low resolution, small images, etc.) and the non-use of biometric techniques.

6.1.2 Feature Extraction

Three types of color features are extracted from torso and leg parts. Their distance with respect to the symmetry axes j_{LRk} is taken into account in order to minimize the effect of pose variations.

6.1.2.1 Weighted Color Histogram

HSV weighted histograms are computed in each of the 4 body parts. The weighting is performed according to the distance of each foreground pixel to j_{LRk} of the region to which it belongs. More precisely, each foreground pixel is weighted by a one-dimensional Gaussian kernel $N(\mu, \sigma)$ where μ is the coordinate of the symmetry axis and σ is a priori set to $J/4$. In this way, pixel values near j_{LRk} have more importance in the final histogram. At the end, four weighted color histograms are obtained (one per part).

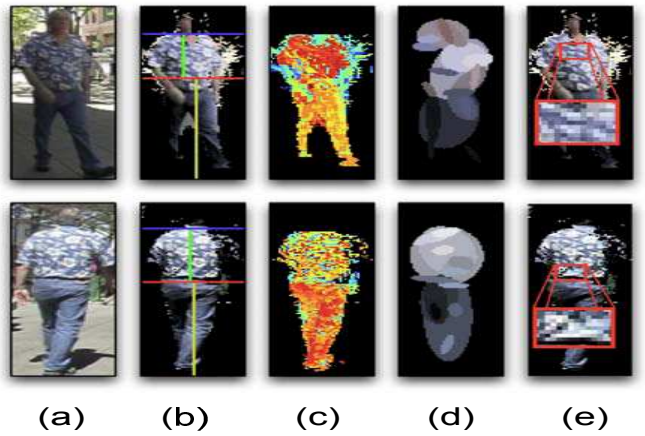


Figure 6.2: Sketch of [Farenzena 2010] approach: a) two instances of the same person; b) x- and y-axes of asymmetry and symmetry, respectively; c) weighted histogram back-projection (brighter pixels mean a more important color), d) Maximally Stable Color Regions; e) Recurrent Highly Structured Patches.

6.1.2.2 Maximally Stable Color Regions (MSCR)

The approach segments the people images in regions with stable colors using the MSCR (Maximally Stable Color Regions) algorithm [Forssén 2007] (see figure 6.2 (d)). The extracted regions are then described by their area, centroid, second moment matrix and average color, forming 9-dimensional patterns.

MSCR algorithm is applied on foreground pixels. In the single-shot case and in order to discard outliers, only MSCRs that lay inside the Gaussian kernel used for color histograms are selected. In the multiple-shot case, the MSCRs coming from the different images are accumulated by employing a Gaussian clustering procedure [Figueiredo 2000], which automatically selects the number of components. The clustering is carried out using the 5-dimensional MSCR sub-pattern composed by the centroid and the average color of the blob. Blobs similar in appearance and position are clustered, since they yield redundant information. This clustering operation allows to capture only the relevant information and keeps low the computational cost of the matching process, where the clustering results are used.

6.1.2.3 Recurrent High-Structured Patches

This new descriptor is proposed by Farenzena et al. [Farenzena 2010] to highlight image patches with texture characteristics that are highly recurrent in the person appearance (see figure 6.3). The extraction of RHSP is performed in three steps:

First, a set of patches p of size $[I/6 \times J/6]$ are extracted randomly on each torso and leg regions independently. These patches are mainly sampled around the j_{LRk}-axes in order to take symmetries into consideration. The more informative patches are selected by thresholding the values of entropy of all the extracted patches. This entropy is computed as

$$H = H_p^R + H_p^G + H_p^B \tag{6.6}$$

where H_p^R, H_p^G, H_p^B are the entropy of red, green and blue channels respectively. The entropy of one channel image is given by:

$$H_p = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{6.7}$$

where P_i is the probability of occurrence of pixel value i and is provided by the corresponding bin of the image histogram.

In the presented approach, patches with H higher than a fixed threshold $\tau_H = 13$ are kept (the authors have fixed this value experimentally)

The second step consists in discarding low recurrent patches. For each patch p, a set of transformations T_i, $i = 1, 2, ..., N_T$ are applied to generate a set of N_T patches p_i, and to obtain an enlarged set $\hat{p} = \{p1, ..., p_{N_T}, p\}$. These transformations T_i consists in rotations along the y central axis of the patch with several angles. The Local Normalized Cross-Correlation (LNCC) is then computed for each patch in $\hat{p}$ by considering only the LNCC value of the local region containing p and not the one of the whole person image. All the N_T+1 LNCC maps (matrix representations of LNCC values for each patch) are then merged together into the average map. Patches containing small values (with respect to a fixed threshold) in this map are discarded.

The last step consists in clustering remaining patches p in order to avoid patches with similar content. This is done using the Gaussian clustering [Figueiredo 2000] on the HSV histogram of the patches, keeping for each final cluster the patch nearest to the cluster's centroid.

In the multi-shot case the candidate RHSPs are accumulated over all frames.

6.1.3 Signature Comparison

To compare images of two persons, a dissimilarity measure is computed between their signatures. The signature consists in the combination of the three extracted features. The dissimilarity measure between a signature I_A and another signature I_B is given by:

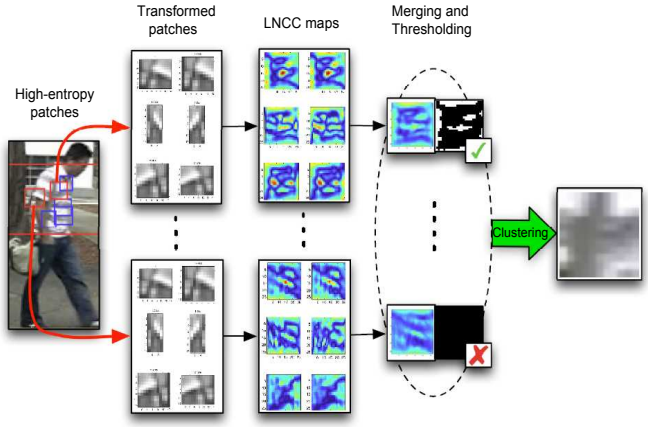


Figure 6.3: [Farenzena 2010] Recurrent High-Structured Patches (RHSP) extraction. The final result of this process is a set of patches (in this case only one) characterizing each body part of the pedestrian.

$$\begin{aligned} d(I_A, I_B) = {} & \beta_{WH}.d_{WH}(WH(I_A), WH(I_B)) + \\ & \beta_{MSCR}.d_{MSCR}(MSCR(I_A), MSCR(I_B)) + \\ & \beta_{RHSP}.d_{RHSP}(RHSP(I_A), RHSP(I_B)) \end{aligned} \tag{6.8}$$

where the $WH(.)$, $MSCR(.)$, and $RHSP(.)$ are the proposed Weighted Histograms, Maximally Stable Color Regions, and Recurrent High-Structured Patches respectively, and β_{WH}, β_{MSCR}, and β_{RHSP} are their normalized weights respectively.

- The distance d_{WH} evaluates the weighted color histograms. The HSV histograms of each part are concatenated, channel by channel, and compared via Bhattacharyya distance. In the multi-shot case, each possible pair of histograms contained in the different signatures are compared, and the obtained lowest distance is selected.

- For d_{MSCR}, in the case of single-shot signature comparison, the final distance between MSCRs is the sum of all minimum distances between all possible pairs of MSCR elements (a, b) (a is an MSCR element from I_A and b is an MSCR element from I_B). This distance is defined by two components: d_y^{ab} that compares the y component of the MSCR centroids, and d_c^{ab} that compares their mean color. In both cases, the comparison is carried out using the Euclidean distance. This results:

$$d_{MSCR} = \sum_{b \in I_B} \min_{a \in I_A} \gamma \cdot d_y^{ab} + (1 - \gamma) \cdot d_c^{ab} \tag{6.9}$$

where γ is a real parameter which takes values between 0 and 1 and is used to take more importance to color position (with respect to the considered axis) than color value or vice versa.

In case of multiple-shot comparison, the first step consists in clustering all the MSCRs of the multiple images of a given person to provide a unique representation of this person in terms of MSCRs, as if it is provided by a single image. This clustering is performed on the images of the same person using the distance d_{MSCR} (eq. 6.9). Once each multiple-shots MSCRs has been clustered and a representation of each person as a unique image MSCR is obtained, the final distance is computed exactly in the same way as in the single-shot case, using eq.(6.9)

- d_{RHSP} is obtained by selecting the best pair of RHSP, one in I_A and one in I_B. The minimum Bhattacharyya distance among the RHSP's HSV histograms is evaluated. This is done independently for each body part, summing all the distances achieved and then normalizing with the number of pairs.

In their experiments, Farenzena et al. [Farenzena 2010] have used the first 100 images of the VIPeR dataset to estimate several parameters values, and have fixed them as follows: $\beta_{WH} = 0.4$, $\beta_{MSCR} = 0.4$, $\beta_{RHSP} = 0.2$ and $\gamma = 0.4$.

6.2 Approach Limitations

6.2.1 Fixed Weights for Each Descriptor

In the initial approach, the weight of each descriptor is fixed using an experimentation on a subset of only 100 images from a unique dataset which is VIPeR (viewpoint invariant pedestrian recognition [Gray 2007]) dataset. These fixed weights are used for all the experiments.

These parameters are "probably" the best ones for VIPeR dataset as they were extracted by experiments on it, but it is clear that each dataset has its own characteristics. More generally, each video surveillance system provides different content for analysis, depending on several conditions: depending on the season/weather, it is observed that people generally wear cloths with uniform/dark colors and poor textures (overcoat for example) in winter or when it is cold while they wear cloths with various bright colors and textures in the summer or when the weather is mild. Indoor and outdoor surveillance systems may also provide different levels of information due to illumination changes. The deployment location of the video surveillance system also may provide different types of content. For example, an airport system will provide more people with

Backpacks or carrying luggages (both can be integrated in the visual signature of people if they are not correctly separated or if they occlude body parts) than a downtown surveillance system. Many other factors may provide different types of content.

In [Meden 2013], the author provides an experimentation on the whole VIPeR dataset to show that only weighted color histograms provide the most significant information and the most important discriminant power of the approach. By testing some combinations of weights for each descriptor, the author demonstrates that MSCR and RHSP does not provide a meaningful improvement. The results of this experiment is provided in the CMC curves of figure 6.4.

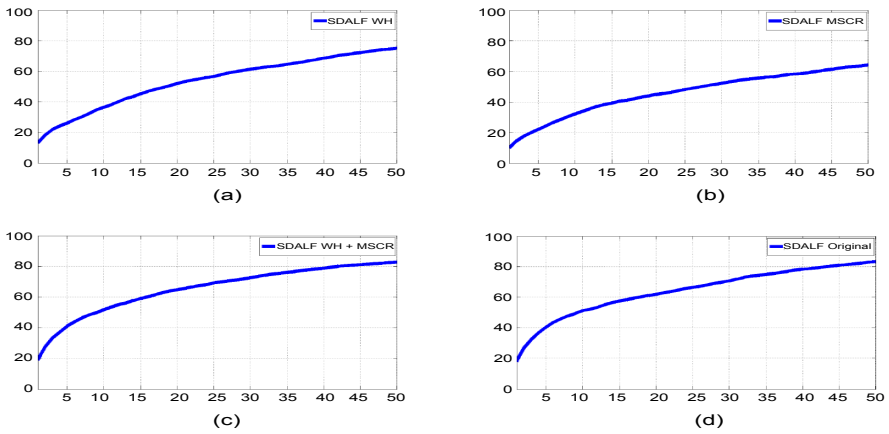


Figure 6.4: Different weighting of SDALF components: (a) Only weighted color histogram (nAUC = 89.11%). (b) only MSCR (nAUC = 83.56%). (c) Weighted color histograms and MSCR with equal weights (nAUC = 91.81%). (d) Original weighting of SDALF like in [Farenzena 2010] (WH = 0.4, MSCR = 0.4, RHSP = 0.2) (nAUC = 91.57%).

CMC (Cumulative Matching Curve) represents the probability to find a correct matching among the r best matching, for r = 1...R (R is the number of sorted matchs). r is the rank of re-identification. In [Bazzani 2012], the nAUC (Normalized Area Under the Curve) value is used to evaluate the re-identification performances. It represents the area under the CMC curve expressed in %. More explanations and details concerning CMC cuves and nAUC are provided in sec. 7.3.1 (Experiments chapter).

Following this experiment, Meden [Meden 2013] uses the following weights: $\beta_{WH} = 1.0$, $\beta_{MSCR} = 0.0$ and $\beta_{RHSP} = 0.0$.

This demonstrate that the original weights provided by [Farenzena 2010] are not necessarily the best on the whole VIPeR dataset, and thereby, may be inappropriate for other datasets, but as we will show in next sections, removing totally the MSCR and RHSP descriptors is not the best way to obtain better re-identification results. An

adaptive weighting system is suitable to deal with the current conditions and to give more importance to the most discriminant descriptor for each situation.

6.2.2 Exclusive Use of Color Information, Without Managing Color Rendering Difference Between Cameras

In the original approach, the three extracted types of feature (Weighted color histograms, MSCRs and RHSPs) use exclusively color information to characterize a person. Even if recurrent high-structured patches (RHSP) are extracted and selected using a complex process based on textures (by selecting patches with a minimal amount of entropy and by performing some transformations to keep the most robust patches to rotation using LNCC maps), the final characterization of each selected patch is performed using a simple color histogram. Adding these histograms to weighted color histograms and MSCR's, which are also color-based descriptors, provide some kind of redundancy and may be not sufficient.

Farenzena et al. use HSV histograms instead of a more specific feature for texture describing. They claim that this is because the RHSP's content is not necessarily a texture, since it exhibits less regularity.

We agree with this last assertion in the sense that effectively, the patches may do not contain enough texture information, but in the case they do, this useful information should be exploited.

For this reason, it is necessary to characterize the patches with both color and texture to provide better results.

More generally, using only color to compute people signature is risky, due to the various situations where this information is not reliable or is poor (the same remark as in the previous paragraph, concerning frequent dark clothes in many situations). For this reason, it is suitable to integrate texture information, which is decorrelated from color information, in the computed signature.

Note that the three initial color features are extracted directly from input images, without managing one of most important color-based signatures issue which is the difference in color acquisition between cameras, i.e. the images of the same object acquired by different cameras may show color dissimilarities. This issue affect greatly the original approach results especially on some dataset like iLids, where many people clothe colors are rendered differently (see in figure 2.26)

6.2.3 Dependency to Orientation

Another important issue of the [Farenzena 2010] approach, which is also a general issue, common to many state of the art approaches, is the dependency of extracted signature to the visible side of people. For example, [Farenzena 2010] approach assumes implicitly that the extracted features from a given person in one camera will be "visible" in the other cameras. This assumption may be verified in several situations especially for weighted color histograms and MSCR if the considered people are wearing clothes with same appearance from all sides, but most of time and depending on the context, this assumption may be not correct. A person wearing an opened jacket with different colors than his/her t-shirt will provide different color histograms and MSCRs depending on whether they are observed from front or from back. Same remark for a person having a backpack or dragging a luggage behind him/her which occlude partially his/her legs. Concerning RHSPs, their are strongly dependent on the visible side of people from which they have been extracted especially as they are more local descriptors than the weighted color histograms and MSCRs.

More generally, we can distinguish three main types of signature computing approaches for multiple-shot case, each of them can be impacted differently with the orientation dependency issue:

- Accumulation and mean/variance modelling: This kind of approach aims at modelling the people appearance by encoding the variations of the observation/features as a "most likely appearance" and an interval of admissible variations for each feature, extracted from the available images of a given person. This kind of approach has the inconvenient to be too permissive and to enlarge the possible variation interval if the accumulation of features is performed without any control process and without any a priori knowledge concerning the provenance of information (visible side of the person). A "large" model may lead to a high rate of incorrect match during the re-identification process, due to its permissivity.

- Filtering and selection of recurrent/constant features: This kind of approach aims at selecting and keeping only features which are stable among all available images of a given person. For example in [Bak 2011], all the images of a given person are resized into the same size and divided on a grid of small subregions. These subregions are characterized by covariance matrices. A mean covariance matrix is computed for each subregion location. Only subregions with salient covariance matrices are kept for the person signature, computed by the proposed reliability measure in [Bak 2011], based on the variance of each region in terms of covariance matrices, or by automatic learning using boosting. This kind of approach is too

restrictive and has the inconvenient to provide a poor signature when the different sides of a given person are significantly different and are all used to compute this kind of signature.

- Use of multiple-shot case as a set of single-shot: This kind of method (mainly the case of [Farenzena 2010] approach) does not combine all the multiple images of a given person in a unique model, but keeps all the single-shot signatures of all people and compares all possible pairs of single-shot signatures from the query person with all signatures of candidate humans, and takes the lowest distance as the one for the given person comparison. This method has the inconvenient to be highly time consuming due to the number of considered person pairs, and the number of used images per person. It is possible to decrease the processing time by sampling images of the same person and by taking a small set only, but without any a priori knowledge concerning the visible side of the person, this sampling may lead to select images of similar content and avoid those representing other situations (other visible sides).

For these reasons, it is necessary to add the people orientation information to the computed signature.

6.2.4 Unreliable Body Subdivision

The symmetry and asymmetry axis estimation proposed in the original approach generally provides correct results when the background subtraction is correct and when the people are wearing t-shirt and pants with different colors. Both of these constraints are not always verified. In the case of bad results of background subtraction or their unavailability (static images, or detected people using a detector instead of background subtraction process) or in the case of uniform cloth color, the torso/legs separation may be incorrect and the asymmetry axis may be shifted up or down with respect to its real position (see figure 6.5 (a)). If the asymmetry axis is positioned differently between each image of the same person in multi-shot case (due to the variation of background subtraction for example), the matching performances may be decreased proportionally to the axis estimation error, due to the weighting process with respect to the position of this axis. We have indicated the torso/legs asymmetry axis manually on a set of images from VIPeR dataset, and we have compared the provided re-identification performances with those provided by automatic axis estimation. We have observed that small errors of positioning (like in figure 6.5 (a)) provide negligible performance difference while more important axis positioning errors lead to decrease the re-identification performances.

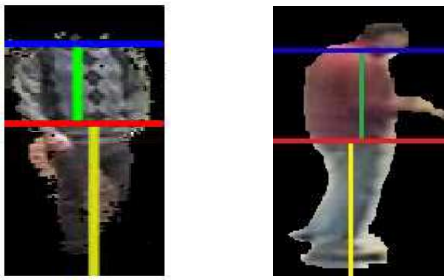


Figure 6.5: Symmetry and asymmetry issues: (a) incorrect torso/legs separation. (b) insignificant symmetry division.

On the other hand, the symmetry division may do not have sense if people are observed from profile. The symmetry axis aims at separating torso and legs into two "similar" and symmetric parts. In the case of profile view (see figure 6.5 (b)), this subdivision may be incorrect and does not provide any useful information. It is then better to perform this second division only if the front or the back side of the person is visible, which leads to the previous mentioned issue, concerning orientation dependence of signatures.

6.3 Proposed improvements

To deal with the previously mentioned issues, we propose the following improvements:

- **Geometrical body subdivision and image alignment:** To deal with the unreliability of the initial body subdivision and to improve information localisation on all images of the same person, we propose to use another method to divide human body, based on statistical dimension of body parts with respect to the whole body height. We also propose a method to align all the images of a given person to extract the corresponding information from the same parts, improving the re-identification process.

- **Color normalization before feature extraction:** To deal with the difference in color rendering between cameras, we propose to use a color normalisation method instead of camera colorimetric calibration approaches, due to the issues and constraints of this last kind of solution, which consists in the non-bijectivity of color projections and to the practical complexity to setup a real deployed camera network system by annotating a sufficient number of people and to perform the learning on all possible pairs of cameras (see 2.3.2.1).

- **RHSP characterisation by color and texture:** The RHSP characterisation by color histograms is replaced by region covariance descriptors, containing both color and texture information, enriching the available information for matching.
- **Use of SIFT descriptors as an additional texture descriptor:** An additional texture information is used in the final signature, without any additional processing cost due to the availability of this information provided by mono-camera tracking process.
- **Use orientation information for visible side classification:** To deal with orientation dependency of visual signatures, each person is represented by a set of sub-signatures, depending on his/her visible side. Each acquired image is assigned to the corresponding sub-signature for update.
- **Use real world positions to filter incoherent/impossible matching:** To decrease the number of candidate for each re-identification query and thereby, the re-identification error rate, a filtering step is performed using real world informations provided by camera calibration information
- **Adaptive weights for each descriptor:** The weights are adapted to each person to provide better re-identification performances. No off-line learning is required to extract selected the weights.

6.3.1 Geometrical Body Subdivision and Image Alignment

As it was mentioned, the proposed symmetry and asymmetry separation is strongly constrained by the quality of background subtraction results and the difference of colors between torso and legs. To be independent from these constraints, and to "save" some processing time which is dedicated to asymmetry/symmetry axes estimation without a guarantee of success, we propose to not perform axis estimations and replace them by a statistical and geometrical body subdivision as in [Huang 2009], by positioning the first asymmetry axis which separates the head from torso at 1/5 of the people image height and the second axis which separates torso from legs at 3/5 of the people image height.

The symmetry axes on the torso and leg parts are kept only if the visible side of the considered person is not known (see sec. 6.3.5). In fact, in addition to the imprecision of their estimation in general case (where background subtraction and people delimitation are not precise), these axes are proposed by [Farenzena 2010] to provide some robustness to the used features against people rotations. Our visible side classification (detailed in sec. 6.3.5) allows to deal with people rotations in a more reliable way. This decision provides another non negligible effect: The weighted color histogram (WH) size

is divided by two (Only whole torso and whole legs histograms are concatenated, unlike for [Farenzena 2010] approach where 4 body part histograms are concatenated). This divide weighted color histograms comparison time by two for two images comparison. By multiplying this saved time per the number of testes pairs of images in multiple-shot images and per the number of considered people, the saved processing time becomes not negligible.

In the following paragraphs, we keep the name of "Weighted Histogram" (WH) in all situations to keep the same feature names as in the baseline approach of [Farenzena 2010], even if in practice, if the visible side is known, the removing of symmetry axes cancels the weighting of pixel colors (pixel colors are weighted with respect to the distance of these pixels to symmetry axes). In this case, we can consider that all pixels have the same weight.

Another issue still persists: most of the time, people are delimited by bounding boxes with acceptable precision, and small errors provide negligible matching errors as mentioned in sec. 6.2.4. However sometimes, the person bounding box may not be precisely computed due to an incorrect background subtraction or an error from the people detector, cropping a part of the body (missing head in figure 6.5 (a)) or adding more background area providing a bad centering of the person within the bounding box. Detecting the images which contain these situations is not a trivial task, and the use of these images for the subdivision of 1/5, 3/5 of the height, and 1/2 of width may be greatly erroneous and it may decrease significantly the re-identification performances. For this reason, we propose also a method to align all the images of a given person provided by the mono-camera algorithm.

This alignment method assumes that most of extracted images of a person are correctly delimited and tries to identify and to remove/readjust those with cropped parts of the person or with additional background.

Given two images I_1 and I_2 of the same person, obtained from the mono-camera tracking algorithm, two cases can be distinguished:

- The two considered images are successive or acquired in a small time interval. In this case, the variations in person size on images, and in the pose are negligible.
- The two considered images are acquired at different moments (low acquisition frame rate or a sparse sampling). In this case, the size (scale) of the person images may be different if one image is acquired when the person is near the camera and the other image when he/she is far from it. The pose of this person may have probably changed too.

Both cases are handled by our alignment algorithm. This algorithm aims at position-

ing I_2 on I_1 so that the body parts (head, shoulders, torso, legs and arms) are aligned as much as possible and with the most similar sizes.

The first step consists in determining the scaling interval. If the images are provided by our mono-camera tracking, which is the case in this work, the camera calibration information and the projection functions detailed in the mono-camera tracking chapter are used to define an interval of scaling factors in which we extract the best factor to apply on I_2 to reach I_1 scale.

It is possible to compute a scaling factor directly by dividing the width and height of I_1 by those of I_2 but due to the motivation of this process, i.e. unreliability of bounding boxes which delimit person images (cropped persons or additional background), the direct scale factor may not be correct. Taking an interval around this factor is possible too and is better than the unique factor obtained by dividing, but the magnitude of this interval cannot be determined without any a priori knowledge. For this reason, using real world information (camera calibration) when it is available (as in our case) is well suitable.

In the case of unavailable real world information (no calibration information available), which is the case of many datasets on which we evaluate our methods too, like VIPeR dataset for example, we use the second method cited in the previous paragraph, that defines a sufficient large interval around the factor computed by dividing width and height of I_2 by those of I_1.

Once the scaling interval is defined, for each scaling factor, the following process is performed:

- Rescale image I_2 with the current computed factor to obtain image $I_2^{'}$

- Downscale both images I_1 and $I_2^{'}$ using an integer factor n by dividing both images on a grid of subregions of $n \times n$, compute the mean color inside each subregion of the grid and assign it to the corresponding pixels of the smaller image. Two corresponding smaller images i_1 and i_2 are obtained.

- Considering the top-left corner of i_1 as the origin, the image i_2 is slided on i_1 browsing many possible positions (x, y), $x \in [x_{min}, x_{max}]$ and $y \in [y_{min}, y_{max}]$, and for each position (x, y) of i_2, a dissimilarity distance between the content of overlapped area of i_1 and i_2 is computed. This global distance of matching is given as the mean of distances between colors of each corresponding pixels in the overlapped area (following figure 6.6):

$$d_{(x,y)}(i_1, i_2) = \frac{1}{n} \sum_{\substack{x \le u < x' \\ y \le v < y'}} d_{L^*a^*b^*}(u, v) \tag{6.10}$$

where n is the number of pixels in the overlapping area and $d_{L^*a^*b^*}(u, v)$ is the distance between colors of two superposed pixels in the overlapped area, computed in $L^*a^*b^*$ color space as follow:

$$d_{L^*a^*b^*}(u, v) = \sqrt{(L_2^* - L_1^*)^2 + (a_2^* - a_1^*)^2 + (b_2^* - b_1^*)^2} \tag{6.11}$$

where (L_1^*, a_1^*, b_1^*) is the color of pixel (u, v) of the image i_1 and (L_2^*, a_2^*, b_2^*) is the color of pixel (u, v) of the image i_2 (in the same referential)

This formula (6.11) is the **CIE76** formula (proposed in 1976). It is the first color-difference formula that has been verified by Lab experimentation. More recent and sophisticated formulas (**CIE94** and **CIEDE2000**) have been proposed to improve the distance computing and to deal with saturated regions, but we keep the first formula (6.11) due to the negligible contribution of the other formulas in our context and to their higher requested processing time (more parameters and mathematical operations per pixel:
$\Delta E_{94}^* = \sqrt{\left(\frac{\Delta L^*}{k_L S_L}\right)^2 + \left(\frac{\Delta C_{ab}^*}{k_C S_C}\right)^2 + \left(\frac{\Delta H_{ab}^*}{k_H S_H}\right)^2}$
and
$\Delta E_{00}^* = \sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + R_T \frac{\Delta C'}{k_C S_C} \frac{\Delta H'}{k_H S_H}}$).
The choice of working in $L^*a^*b^*$ color space is motivated by the fact that this color space provides the most similar color distribution as human perception and allows better color comparison. The non-linear relations for L*, a*, and b* are intended to mimic the non-linear response of the eye. Furthermore, uniform changes of components in the L*a*b* color space correspond to uniform changes in perceived color, so the relative perceptual differences between any two colors in L*a*b* can be approximated by treating each color as a point in a three-dimensional space (with three components: L*, a*, b*) and taking the Euclidean distance between them ([Jain 1989]).

- Once all dissimilarity distances are computed for all possible positions of i_2 in the defined intervals, the coordinates $p_{(1,2)}(x, y)$ which corresponds to the lowest distance are taken as the ones providing the best alignment.

- A reverse computing is performed by multiplying the found coordinates $p_{(1,2)}(x, y)$ by n to obtain the superposing coordinates of the original images.

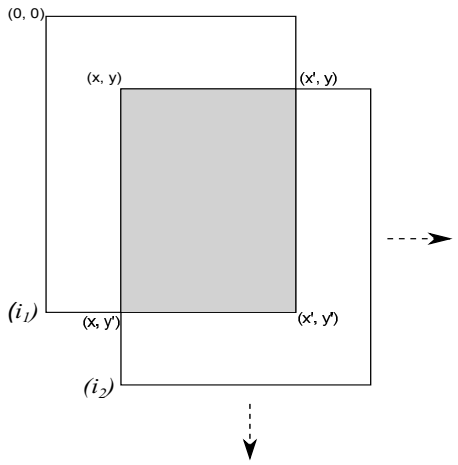


Figure 6.6: Images alignment

For a set of q images $\{I_1, I_2, ...I_q\}$ of a given person, which is the general case, this process is performed for each pair of successive images I_m, I_{m+1} providing $q-1$ superposition points $p_{(m,m+1)}(x, y)$. These coordinates $p_{(m,m+1)}(x, y)$ are relative to different origins which are the top-left corner of the first image of the considered pair of images.

Once the process is achieved, all these coordinates are reported to an absolute origin which is the top-left corner of the first image I_1 of the considered set. This is done easily by a simple summation as long as the pairs are processed in successive order.

With the assumption that a sufficient number of images are correctly delimiting the person, we have observed than these images have very close superposing coordinates, which are null or very close to the origin if the first image delimits well the person.

The largest group of closest superposed images is considered as the correct delimitation of the person. All the other images which are slightly shifted/rescaled in comparison with the largest group of superposed images are kept after being adjusted. The other images which have an important shift (i.e. important crop) are dropped and not considered for the signature computation due to the lack of useful information.

Note that the downscale step is performed to speed up the alignment process. It is possible to apply this algorithm without downscaling, but the required processing time for testing all possible superposing positions after downscaling by a factor n is $\approx n^2$ lower. This factor has to be multiplied by the number of image pairs to align per person. We can then deduct the importance of saved processing time.

In our method, n depends on the size of people images. For example, we use $n = 3$ for small images like for CAVIAR dataset while we use $n = 5$ or $n = 7$ for large people images like for iLids dataset. For the shifting interval, by considering W_1 and H_1 respectively the width and the height of the image i_1 (of each pair of images), we use:

$x_{min} = -W_1/2$
$x_{max} = W_1/2$
$y_{min} = -H_1/2$
$y_{max} = H_1/2$

Some examples of the image alignment results are shown in figure 6.7.

6.3.2 Color Normalization Before Feature Extraction

As mentioned before, the color rendering difference between cameras, due to sensor sensibility difference or to external conditions (camera point of view/orientation, illumination conditions, etc.) is an important issue for visual signature comparison as long as color is an important information for this purpose.

Due to the issues and constraints of colorimetric calibration approaches, discussed at the end of section 2.3.2.1, related to the non-bijectivity of transfer functions and to the complexity to apply this kind of method in a large scale video-surveillance system with many cameras, we prefer to use a color normalization method.

Like in [Bak 2011], we use histogram equalization [Finlayson 2005] method. This method supposes that the rank of colors are preserved during illumination changes. The rank measured for a level "i" of a channel "k" is given by:

$$M_k(i) = \sum_{u=0}^{i} H_k(u) / \sum_{u=0}^{Nb} H_k(u) \tag{6.12}$$

where $H_k()$ is the histogram of the channel k and Nb its bins number.

Histogram equalisation stretches a range of histogram to be as close as possible to a uniform histogram. It is applied to each color channel (RGB) to maximize their entropy and obtain an invariant image.

6.3.3 RHSP Characterisation by Color and Texture

To deal with the exclusive use of color information for the signature computation, two improvements are proposed. The first one consists in the modification of the RHSP characterization method, and the second one, which is detailed in the next section, consists in the addition of another texture-based feature to the signature.

As mentioned before, the extraction method of RHSP patches, described in 6.1.2.3, provides a set of recurrent patches, which are also robust to rotations. Unfortunately, the final characterization of these patches is performed by simple color histograms, loosing the information of color repartition inside the patchs and does not encode any texture information, despite the use of LNCC in the selection process.

Figure 6.7: Image alignment process illustration and results on iLids-AA dataset samples. In the first row, the alignment process is detailed: (a) 8 sampled images of the same person obtained by automatic people detection and tracking. The person is not delimited and centred correctly in images (various shifting). (b) Samples of alignment of successive pairs of images. (c) The final alignment results. (d) Some other results of people images alignments. In this figures, red bounding box corresponds to the first image of each set of images (defining the origin of coordinate system), and the blue bounding boxes are those of the remaining images of each set. Image transparencies have been modified to show the alignment results.

We keep the same method for patches selection, but we propose to replace color histograms by region covariance descriptors (see sec. 4.1), containing both color and texture information. We use the following feature vector to construct our 7×7 covariance descriptors:

$$\left[x \;\; y \;\; R \;\; G \;\; B \;\; \sqrt{I_x^2 + I_y^2} \;\; \arctan\frac{|I_x|}{|I_y|} \right]^T \tag{6.13}$$

where:
x and y are the pixel coordinates,
R, G and B are respectively the values of red, green and blue channels,
$\sqrt{I_x^2 + I_y^2}$ and $\arctan\frac{|I_x|}{|I_y|}$ are respectively the magnitude and the orientation of the gradient (the edge) computed on the green channel (to save conversion processing time).

Using these features provide the needed information: the localisation of both color and texture information (x and y coordinates), the color information (R, G, and B) and the texture information ($\sqrt{I_x^2 + I_y^2}$ and $\arctan\frac{|I_x|}{|I_y|}$).

Note that contrary to the use of covariance descriptor for people detection (chapter 4), where the processing time is relatively high, the use of this descriptor here does not increase the processing time. This is due to the fact that no mean covariance is required (and thereby, no iterative gradient descent computations), but only a simple covariance matrix computing for each patch, which can be speeded up using integral images, as explained in the chapter 4.

6.3.4 Use of SIFT Features as an Additional Texture Descriptor

To enhance the final signature by texture information, the SIFT features which have been used in mono-camera tracking are added to the final signature. The last known state (coordinates and descriptors) of all SIFT features of a given person are stored and used to provide a partial matching score which is weighted and summed to eq. 6.8). The new signature matching formula becomes:

$$\begin{aligned} d(I_A, I_B) = {} & \beta_{WH}.d_{WH}(WH(I_A), WH(I_B)) + \\ & \beta_{MSCR}.d_{MSCR}(MSCR(I_A), MSCR(I_B)) + \\ & \beta_{RHSP}.d_{RHSP}(RHSP(I_A), RHSP(I_B)) + \\ & \beta_{SIFT}.d_{SIFT}(SIFT(I_A), SIFT(I_B)) \end{aligned} \tag{6.14}$$

where the $SIFT(.)$ is the partial signature consisting of SIFT features, and β_{SIFT} is its normalized weight in the final signature.

The distance d_{SIFT} evaluates the SIFT features similarity. Given two persons A and B to compare, each of them has its set of SIFT features provided by mono-camera tracking algorithm.

As described in sec 5.2.2, each person image is divided into a grid of subregions and each subregion contains a constant number of SIFT features. Thanks to the images alignment algorithm we use (sec 6.3.1), we can assume that these two persons are well delimited by their bounding box. It is then possible to resize their images to the same dimensions and thereby to obtain aligned grids of subregions and to compute the corresponding SIFT feature coordinates in the new images size.

Each SIFT feature $f_i(A)$ from A is compared to a set of SIFT features $\{f_j(B)\}_{j:1..n_B}$ from B. This set of features from B consists of the n_B SIFT features which are inside the same grid cell (subregion) than $f_i(A)$ and those inside adjacent cells (subregions) to take in account possible shifting of features located near subregions borders. We also compute a SIFT descriptor on B image at the same location than $f_i(A)$ even if is not a detected SIFT point, and compare it with $f_i(A)$. This last point has the interest to ensure one point at least for comparison, so each SIFT feature from A provides a matching distance.

The comparison between two SIFT features is performed using an Euclidean distance like during mono-camera tracking. The smallest distance is taken as the one of $f_i(A)$ with its corresponding feature in B. The final distance d_{SIFT} is taken as the mean of all the smallest distances between all $\{f_i(A)\}_{i:1..n_A}$ and their corresponding features in B.

Concerning β_{SIFT} and as it was indicated in sec 6.2.1, this weight (as all the other) is no longer constant in our approach and is fixed according to the available information as it is described in the next section (sec 6.3.7).

Note that due to the local nature of both RHSP and SIFT features, their corresponding signatures are strongly dependent on the visible side of the person. This issue is managed by the use sub-signatures per person, detailed in sec 6.3.5

6.3.5 Use Orientation Information for Visible Side Classification

As it was mentioned before, the orientation change of a given person is an important issue for re-identification task. Both global and local features may be affected with different levels.

It is possible to use face detection to identify whether an observed person is in front of the camera or not, but this solution presents the following issues: First, face detection requires additional processing time, slowing down the whole process. Then, the detection performances are strongly dependent on image resolution and face sizes, which are not appropriate with most of large area surveillance systems. Finally, if no face is de-

tected, that mean that the person is not observed on his frontal side. It can be observed from behind or in profile. A binary classification of visible side (frontal side / not frontal side) is not sufficient.

To make the visual signature more discriminant, we have decided to consider 8 possible classes for visible sides for each person (see figure 6.10(b)). Each visible side correspond to a walking direction. These 8 side classes are: the 4 main sides which are the front side (S), the back side (N), the left profile (W) and the right profile (E), in addition to the 4 intermediate sides which are the front-left profile (SW), the front-right profile (SE), the back-left side (NW) and the back-right side (NE).

The availability of camera calibration information and the real world information concerning people movements, provided by mono-camera tracking algorithm allow us to obtain more detailed information concerning the visible side of each person, assuming that people move forward.

The used calibration tool (figure 5.2) allows to position the the coordinate system as desired on the ground floor. If the coordinate system is not positioned manually, a default position is assigned by the calibration tool as follow (see figure 6.8):

- The origin "o" is positioned at the vertical projection of the optical center of the camera on the ground floor.
- The "Y" axis is perpendicular to the ground floor and is directed upwards and thus it passes through the optical center of the camera (hence, the term "Y-top" for this kind of calibration).
- The "X" axis is the projection of the optical axis of the camera on the ground floor, and is oriented in the camera view direction.
- The "Z" axis is defined as perpendicular to the plan defined by the two axes "X" and "Y". The direction of "Z" axis is defined using the standard "right-hand rule" (see figure 6.9).

We use this default positioning and orientation of the coordinate system. If another calibration tool is used with a different coordinate system configuration, obtained extrinsic matrix has to be transformed using the necessary translation and rotation matrices to reach the desired configuration of the coordinate system.

The mono-camera tracking algorithm provides the trajectory on the ground floor of each tracked person. The trajectory of a given tracked person consists in the set of its localization coordinates $\{(X_i, 0, Z_i)\}_{i:1...N}$ on the ground floor, where N is the number of frames where the tracked person is tracked, obtained using the "image to world" projection process (explained in sec 5.1.1).

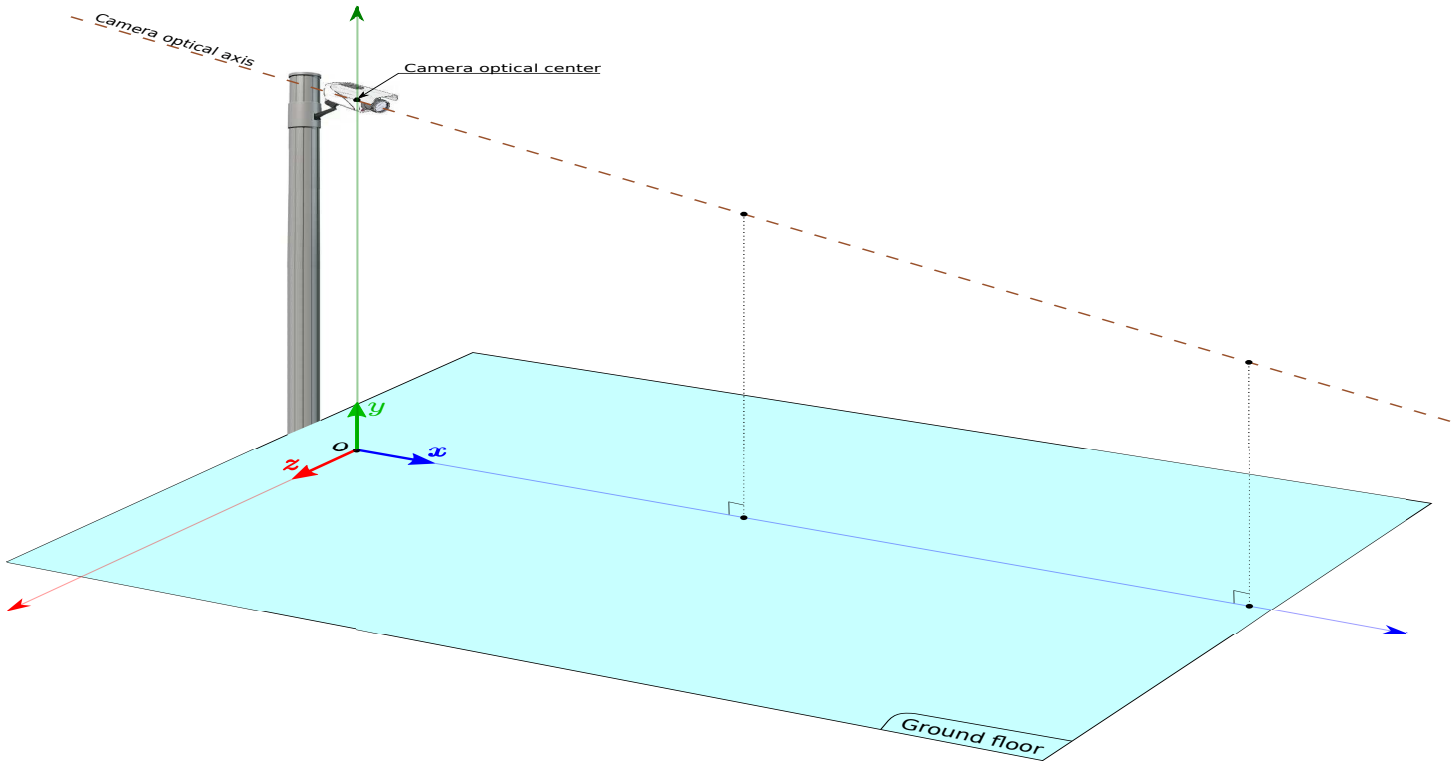


Figure 6.8: Used (default) camera calibration coordinate system.

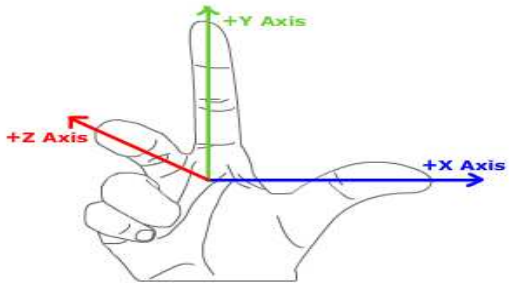


Figure 6.9: Right-hand rule for Cartesian 3D coordinate system.

By ignoring Y coordinate (which is always null), we can consider the trajectory of a given person in the corresponding two dimensional plan to the ground floor (see figure 6.10 (a)).

For a given person, his/her trajectory points are used to compute a regression line. The slope of this regression allows to classify the visible side into one of the 8 defined classes (see figure 6.10 (b)). Unfortunately, people trajectories are not always straight, so the regression line of the whole points of the trajectory may not be representative of all observed visible sides. The trajectory of “Person 4” in the figure 6.10 (a) shows that 3

different sides have been observed successively (left side, front-left side and front side) while the global regression line correspond to a unique front-left side.

To deal with this frequent issue (no straight trajectories), the trajectories are divided into subsets of "n" consecutive location points. We consider that the change in visible side may not be important in a walking interval of 2 meters. Knowing that the average human walking speed is about 5km/h ([TranSafety 1997]), a person need approximately 1.44 seconds to walk 2 meters. Depending on the acquisition frame-rate, it is then possible to define the number "n" of successive location points which are traveled by a person in this time. In our case, most our used sequences are acquired at 8 fps by deployed systems of Digital Barriers, so the corresponding position points is 11.52 which is rounded to $n = 12$ points. This value has to be adjusted according to the average walking/running speed in the monitored area and to the acquisition frame-rate.

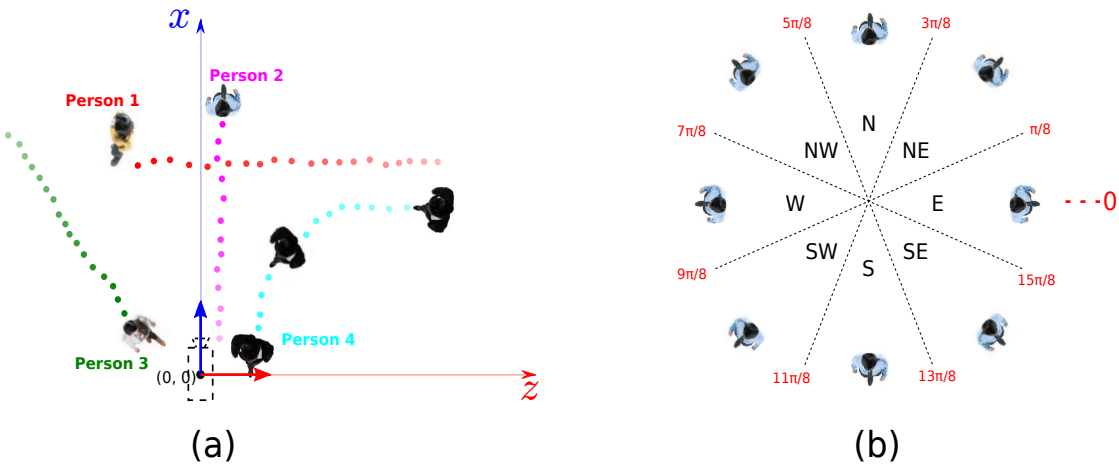


Figure 6.10: Visible side classification into 8 sub-classes according to the walking direction. (a) Example of trajectories in the ground floor projection plan. (b) The 8 classes of visible sides

To increase the precision of the visible side classification, we do not assign the current person image to it's visible side class directly using the regression line on the last "n" position. A buffer of "$n/2$" images (6 in our case) is used to assign the image to its class with a delay of "$n/2$" images. This policy allows to take the "$n/2$" previous locations and the "$n/2$" next locations of a given image to decide in which class it belongs. Otherwise, if the current image is directly classified using the "n" last locations (including the current one as the last one), there is a risk that the current image correspond to the end of orientation change (see the last position of the "Person 4" in figure 6.10 (a): the visible side is the front one while it may be classified as a front-left side if the orientation change is rough).

Now it is possible to estimate the visible side of a given person on each acquired

image, his/her global signature is replaced by a set of 8 sub-signatures, one per class of visible side. One or several visible side classes may have null signatures due to the fact that the person has never been observed by this side(s).

To compare the signatures of two persons A and B, three cases can be distinguished:

- The two persons share some classes with non-null sub-signatures (i,e. they have been observed under some similar sides).
- The two persons do not share any similar classes of non-null sub-signatures, but some of their non-null sub-signatures belongs to adjacent classes (for example, the person A has a "N" sub-signature while the person B does not have a a "N" sub-signature but have a "NE" or "NW" one).
- The two persons do not share any similar classes of non-null sub-signatures, and any adjacent classes of non-null sub-signatures.

In the first case, where at least one non-null sub-signature class exists for both persons, the dissimilarity measure between the signature S_A of A and the signature S_B of B is given by:

$$d(S_A, S_B) = \frac{1}{n} \sum_{i=1}^{n} d(SS_A^{(i)}, SS_B^{(i)}) \tag{6.15}$$

where n is the number of common classes with non-null sub-signatures, $SS^{(i)}$ is the sub-signature of the class i, and $d(SS_A^{(i)}, SS_B^{(i)})$ is the dissimilarity distance between two sub-signatures, computed using eq. (6.14) as for usual signatures before.

In the second case, where only some adjacent classes of sub-signatures are shared between the two persons (no common classes), the dissimilarity measure between the signatures S_A and S_B is given by:

$$d(S_A, S_B) = \frac{1}{n} \sum_{i=1}^{n} d(SS_A^{(i)}, SS_B^{(j)}) \tag{6.16}$$

where n is the number of pairs of adjacent classes with non-null sub-signatures, $SS^{(i)}$ and $SS^{(i)}$ are respectively the sub-signatures of the adjacent classes i and j, and $d(SS_A^{(i)}, SS_B^{(j)})$ is the dissimilarity distance between two sub-signatures, computed using eq. (6.14) as for usual signatures before.

Finally, in the last case where neither common classes nor adjacent ones are shared, the dissimilarity distance between the two persons is computed as in the initial approach, by considering all the sub-signatures as a unique one, but using the new formula 6.14

including SIFT features signature. In this case, the result is similar to the one of the initial approach (not worsen).

This subdivision into 8 classes of visible sides has the following three main advantages:

- In the general case, when the computed visual signature consists of a large model encoding all the possible observations from all acquired images of a given person, or consists of a selection of salient features which are present in all person images, this sub-signatures subdivision allows the extraction of information from the same sides and thereby, avoids too permissive or too restrictive signatures obtained from important variations in the appearance of each person side.

- In our case, where the comparison between two persons consists in keeping the minimum distance of all tested pairs for *WH* and RHSP, the sub-signatures decrease the number of pairs to test by focusing on smaller subset of images belonging to the selected classes.

- It allows a better feature weighting (β_{WH}, β_{MSCR}, β_{RHSP}, β_{SIFT}) in the final dissimilarity measure computation (See next section). Due to the local nature of SIFT and RHSP features, their weight may be increased, decreased or totally avoided depending on their provenance (if they are not extracted from common sides).

Two important remarks have to be taken into account here:

- If the number of available images for a given person is high, using all these images may provides the best results but due to processing time requirements, we use only a smaller subset of these images. Our criteria for images sampling is mainly based on the visible side classification: For each not empty class, we select a constant number "n" of images at constant intervals of the trajectory which belongs to this class. In our experiments, we select the value of n according to the number of non-empty classes, the aim being to take 8 to 16 images at most. The selection is done by taking images at approximately constant intervals of position in the same visible side class. If no visible side classification is available, the 8 images are sampled with the same method as they belongs to the same class.

- After a full computing of MSCR features for a given image (generally the first one of a class of visible side), the MSCR computation on the next images of the same visible side class is performed faster. For each new image, we initialize MSCR detector with the optimal parameters (number of clusters to find, initial mean colors, initial color region centers and areas) according to the detected MSCRs on

the previous image, assuming that images of the same sides provides closer MSCRs. This speeds up MSCR computing process. In the case this visible side classification is not available, the same process is applied on all the acquired images.

The improvements provided by this contribution are evaluated and highlighted in section 7.3.3.2. It consists on an experimentation and performance comparison on an extracted dataset which contains mainly people who changes significantly their visible side (by rotation) with respect to the camera.

6.3.6 Use Real World Positions/Velocity to Filter and Weight Matching

Depending on the availability of the monitored area plan (containing the delimitation of observed area per camera and the distances between them) and/or the calibration of the deployed cameras in this are in a common coordinate system, the re-identification process can be strongly improved in terms of performances and processing time.

In our study, we use a second tool for camera network calibration, developed by Digital Barriers France for the ViCoMo project, providing calibration matrices for each camera, but reported to a unique coordinate system.

This part of improvements has been tester only on ViCoMo project sequences and has proven his effectiveness. The ViCoMo project video sequences were acquired in Eindhoven airport by 10 cameras, providing both overlapping and non-overlapping fields of view situations. Unfortunately, It is not possible to evaluate it on the most used benchmarking re-identification datasets (iLids, TrecVid, Viper) due a lack of time and to the fact that dataset images are provided as detected and cropped persons, without any real world information. We are planing to evaluate this part on large video sequences like TrecVid ones in our feature work when more time will be available.

We can distinguish two different situations concerning the re-identification between two cameras: The first case concerns cameras with overlapped fields of view. The second one concerns the case of the cameras do not have overlapped fields of view.

- **Cameras with overlapped fields of view:** in this case, the re-identification of a person from the first camera to the second one is performed when this person is observed by the two cameras simultaneously (i,e. when this person is on the overlapping area).

  Due to the possible imprecision of the detection bounding box (even provided by background subtraction or by people detector), the same person may provide slightly different coordinates on the ground floor between the two cameras. This is not an issue if the person is alone in a relatively small surrounding perimeter,

but if more than one person is near the provided locations (which is the case in the processed ViCoMo sequences), the re-identification became non-trivial.

For this reason, a circular perimeter with a sufficient radius "r" is centered on the person position from the first camera. All persons who are out of this perimeter are not considered as candidates. The persons inside this perimeter are tested as candidate, by comparing their visual signatures with the query person one. In our experiments on ViCoMo video sequences, we use $r = 2m$.

The visual signatures comparison is performed as described in previous sections. The dissimilarity measure between two signatures is weighted by the normalized distance (on the ground floor) of the corresponding candidate location to the one of the query person.

$$d(A,B) = d(I_A, I_B).\frac{\sqrt{(X_A - X_B)^2 + ((Z_A - Z_B)^2}}{r} \tag{6.17}$$

where $d(A,B)$ is the final distance used to perform the re-identification task and $d(I_A, I_B)$ is the dissimilarity distance between visual signatures, provided by eq. (6.14).

Note that we are assigning the same importance to the visual signature comparison and to the real world distance. In future work, we are planning to find a better weight for real world distance, taking in account the calibration precision and the number of persons inside the perimeter (the more persons inside the perimeter are, the more the risque of imprecise real world distance computation is, due to increased risk of incorrect projections)

- **Cameras without overlapped fields of view:** if the monitored area plan is available, the velocity of tracked people and their location at the re-identification moment are used to remove all candidates who can not satisfy the spatio-temporal constraints: A person cannot be at two different locations at the same time. A person cannot travail a given distance in a given time if this displacement requires a superhuman velocity. We have used the maximum possible velocity as 44.72 km/h which was the maximum observed human velocity (reached by Usain Bolt in a 100 meters sprint). We know that we can decrease this maximum velocity because normal people, in usual monitored areas have less important velocity, but despite this high value, a large percentage of incoherent candidates have been filtered in our tests on ViCoMo sequences.

  For future work, we are interested by finding a reliable way to learn automatically the possible trajectories and the distances between the covered areas of a camera

network. A first idea consists in assuming that people move with constant velocities and using the elapsed time between the exit from one camera view and the appearance in another one, it is possible to estimate a traveled distance. To avoid cases where a person stops for few time between the two cameras, or he/she takes a small/large detour, it may be necessary to use a large set of manually annotated trajectory samples and to perform some clustering to eliminate outliers.

6.3.7 Adaptive Weights for Each Descriptor

As mentioned in sec 6.2.1, Farenzena et al. [Farenzena 2010] use fixed weights for all the experiments. These weights are extracted experimentally as the best ones for a subset of 100 persons from VIPeR dataset. This approaches is limited due to the possible diversity of situations (indoor/outdoor, weather, illumination, location, visible side of people, etc.) and their provided type of information (poor/rich amount of color/texture).

We propose to use adaptive weights according to the kind of available information and to the visible side of each person, and this without any necessary offline learning, allowing to use this method easily in live surveillance system.

The most important point in our weighting method is that the used weights are not specific for a whole dataset or for all viewed persons in a given surveillance system, but the are specific for each query person independently, and more precisely, they are specific to each considered visible side information. The weights used to re-identify a person A may probably be different from those which are used to re-identify a person B even both persons are belonging to the same dataset or are observed by the same camera.

This policy has the effect of providing heterogeneous dissimilarity distances for different query persons, but it does not constitute an issue for the re-identification task as long as the dissimilarity distances are homogeneous for a given query person, i.e, the query person is compared to all the candidates using the same weights, providing coherent ranking. The lowest distance correspond to the most likely correspondence.

The weights are assigned to a given person depending on the his/her visible side(s), the importance and the discriminative power of the visual information this person provides. SIFT and RHSP features being local features, their weights vary according to the compared visible sides. On the other hand, the more his/her appearance is rich/discriminant in terms of colors, the higher β_{WH} and β_{MSCR} are. The same remark concerns the richness power of his/her appearance in terms of texture and the importance of β_{RHSP} and β_{SIFT}.

The following paragraphs explain how a given person appearance is considered as rich or poor in terms of color/texture and how, according to this decision, the different

weights are assigned.

6.3.7.1 Color/Texture Importance Measures

Color:

Each time a new image of a person is acquired by a given camera, the different features (WH, MSCR, RHSP) are computed on this image and used to update the person signature as described before. At the same time, we extract a Hue histogram of this person image, we assign it to the considered person as additional information (it is not used for signature comparison) and we add it to a global histogram assigned to the camera network (no additional image browsing is needed, the WH bins which belong to the same hue value are summed). The camera network has its own normalized Hue histogram representing the frequency of all observed colors on all observed people.

When a given person re-identification is required, the distance between the hue histogram of each candidate person and the global hue histogram (of the camera network) is computed. The max distance is kept as $Dist_{color}(max)$.

The distance between the hue histogram of the person p to re-identify and the global hue histogram is also computed and noted as $Dist_{color}(p)$. Our color importance estimation is based on our assumption that the higher $Dist_{color}(p)$ is, the more informative the color of this person is. This is due to the fact that the global hue histogram of the camera network is representing a mean of observed colors. If $Dist_{color}(p)$ is high, it means that this person has a high probability to be separated from most of people thanks to its colors.

For the considered person p and to quantify this importance as an importance score, we use the following formula:

$$Score_{color}(p) = \frac{Dist_{color}(p)}{Dist_{color}(max)} \tag{6.18}$$

$Score_{color}(p)$ values are in $[0, 1]$. We are aware that computing an importance score with respect to the max distance is not the best way to obtain an absolute information about the richness of color and its discriminative power, especially if $Dist_{color}(max)$ is too low, but this method provides good results as it is demonstrated in the evaluation chapter (chapter 7). We believe that the main reason for this is that when $Dist_{color}(max)$ is too low, it corresponds to the case where all of people have quite similar colors. In large scale video-surveillance systems or large people datasets, this case is unlikely to occur, since if only one person has significantly different colors, $Dist_{color}(max)$ will be high enough. Even assuming that no person has significantly different colors compared to all the other people, this case happen generally with dark

colors. In this case, the texture features (RHSP and SIFT) are generally impacted too (low textures because dark cloths), and since feature importance weighting is a relative one (Color VS. texture), the computed weights still have sens even if both color and textures are not discriminative enough (in similar proportions).

The Weighted histogram and MSCR features are not fully redundant and have the same importance in our approach since they are complementary and provide different contribution to the visual signature even if both are color-based features. MSCR encodes the spacial information of color distribution and some discriminative color region shapes and orientations (color stripes, circles, etc.) but are more sensitive to deformations (region centroid and orientations may vary) while color histograms are more robust to deformations but does not encore any spacial information.

RHSP:

A similar process as for color is performed. Each time a new image of a person p is acquired by a given camera, the different features (WH, MSCR, RHSP) are computed on this image and used to update the person signature as described before. The mean entropy of the selected RHSP patches (eq. 6.6) for this person is computed and assigned to him/her as additional information (it is not used for signature comparison). This value is noted $Mean_{RHSP}(p)$.

The max value of of all computed $Mean_{RHSP}(p_i)$, noted $Mean_{RHSP}(max)$, is assigned to the camera network as the corresponding value to the most textured person image.

Our RHSP importance estimation is based on our assumption that the higher $Mean_{RHSP}(p)$ is, the more informative the texture of this person is (both gray values and color textures since we use region covariance descriptor with both information to characterise RHSPs). If $Mean_{RHSP}(p)$ is high, it means that this person has a high probability to be separated from most of people thanks to its texture.

For the considered person p and to quantify this importance as an importance score, we use the following formula:

$$Score_{RHSP}(p) = \frac{Mean_{RHSP}(p)}{Mean_{RHSP}(max)} \tag{6.19}$$

$Score_{RHSP}(p)$ values are in $[0, 1]$.

SIFT:

For SFIT features, two cases can be distinguished: If the SIFT features of the considered person p are provided by mono-camera tracking algorithm, we use the mean of their reliability measures, provided also by mono-camera tracking (eq. 5.14) as SIFT

importance score $Score_{SIFT}(p)$. $Score_{SIFT}(p)$ values are in $[0,1]$ since each SIFT feature reliability value is in $[0,1]$ (eq. 5.14). Otherwise, if SIFT features are detected directly on the last image of the person like for some evaluation datasets (see sec. 7.3) (due to the unavailability of tracking information), we use $Score_{SIFT}(p) = Score_{RHSP}(p)$. SIFT features being texture descriptors, we assume that the importance may be quite similar.

6.3.7.2 Feauture Weighting

Now the importance of each information type measured for a given person p, the used features (WH, MSCR, RHSP, SIFT) have the following intermediate normalized weights:

$$w_{WH}(p) = \frac{2 \cdot Score_{color}(p)}{2 \cdot Score_{color}(p) + Score_{RHSP}(p) + Score_{SIFT}(p)} \tag{6.20}$$

$$w_{MSCR}(p) = \frac{2 \cdot Score_{color}(p)}{2 \cdot Score_{color}(p) + Score_{RHSP}(p) + Score_{SIFT}(p)} \tag{6.21}$$

$$w_{RHSP}(p) = \frac{Score_{RHSP}(p)}{2 \cdot Score_{color}(p) + Score_{RHSP}(p) + Score_{SIFT}(p)} \tag{6.22}$$

$$w_{SIFT}(p) = \frac{Score_{SIFT}(p)}{2 \cdot Score_{color}(p) + Score_{RHSP}(p) + Score_{SIFT}(p)} \tag{6.23}$$

with $w_{WH}(p) + w_{MSCR}(p) + w_{RHSP}(p) + w_{SIFTH}(p) = 1.0$.

The final weights β_{WH}, β_{MSCR}, β_{RHSP} and β_{SIFT} of the used features for signature comparison are given by:

$$\beta_{WH}(p) = \alpha_{color} \cdot w_{WH}(p) \tag{6.24}$$

$$\beta_{MCSR}(p) = \alpha_{color} \cdot w_{MCSR}(p) \tag{6.25}$$

$$\beta_{RHSP}(p) = \alpha_{RHSP} \cdot w_{RHSP}(p) \tag{6.26}$$

$$\beta_{SIFT}(p) = \alpha_{SIFT} \cdot w_{SIFT}(p) \tag{6.27}$$

where α_{color}, α_{RHSP} and α_{SIFT} are the visible side classification coefficients. It means that depending on the availability or not of the visible side classification, and to the compared visible side classes if they are available, the feature weights are different.

In the case of availability of visible side classifications and depending to which of the three mentioned cases in sec 6.3.5 the re-identification query belongs, the weights are assigned as follow:

- **Comparison of the same visible side classes:**
  In this case, all the features have the same importance from visible side point of view. The difference between their importance is then exclusively dictated by the intermediate weights (eq. 6.20, 6.21, 6.22 and 6.23) which are directly related to the information importance measures. The coefficients are then: $\alpha_{\text{color}} = 1.0$, $\alpha_{\text{RHSP}} = 1.0$ and $\alpha_{\text{SIFT}} = 1.0$

- **Comparison of the adjacent visible side classes:**
  In this case, Color being more global descriptors than RHSP and SIFT (which are local descriptors), the color features (WH and MSCR) take more important weights. We have decided to keep the same proportion between these two kinds of information (color and texture) as in the baseline approach (sec. 6.1.3), i.e. 80% of weight to color based features (WH and MSCRs) and 20% of weight to texture based feature (RHSPs, even if their characterisation was exclusively done by color, their selection are texture-based approach).

  We assign the coefficient as follow: $\alpha_{\text{color}} = 0.4$, $\alpha_{\text{RHSP}} = 0.05$ and $\alpha_{\text{SIFT}} = 0.15$, with $\alpha_{\text{RHSP}} + \alpha_{\text{SIFT}} = 0.2$.

  We assign a more important coefficient to SIFT features in comparison with RHSP due to their good invariance with respect to affine transformation (object rotation) ([Lowe 2004]) in comparison with covariance matrices. In fact, both SIFT features and RHSP patches may still be visible on images of adjacent classes of visible sides but with some affine transformation.

  Note that in this case, another normalization step, as in equations 6.20, 6.21, 6.22 and 6.23, is required for β_{WH}, β_{MSCR}, β_{RHSP} and β_{SIFT} weights since their sum does not equal 1.0 in general case.

- **Comparison of signatures without common or adjacent visible side classes:**
  In this case, we completely avoid the use of SIFT features and RHSP for visual signature comparison, since they are local descriptors and then, the same features are not visible on the two considered signature, by setting their coefficient as $\alpha_{\text{RHSP}} = 0$ and $\alpha_{\text{SIFT}} = 0$. Color coefficient is set as $\alpha_{\text{color}} = 1.0$, and as for the previous case, β_{WH}, β_{MSCR} are normalized again to have $\beta_{WH} + \beta_{\text{MSCR}} = 1.0$.

In the case of unavailability of visible side classification, we do not have any information to decide if local features (SIFT and RHSP) have to be considered or not, and

with which importance. We believe that the middle case (using these features but with a lower importance) is a good compromise to exploit their effectiveness if the compared signatures are acquired from same or close visible sides, and to do not alter the comparison process greatly if the visible sides are neither common nor adjacent, thanks to the low coefficient these local features have.

6.4 Conclusion

We have proposed a context-aware and appearance based approach for people re-identification through video camera network. This approach is fully on-line processing and satisfies the genericity and easy deployment on large scale video systems constrains. Our approach is based on a state of the art one ([Farenzena 2010]). The main issues of the baseline approach have been identified and some improvements have been proposed to deal with these issues. Some other improvements have been added to increase the efficiency of the final approach.

From visual signature efficiency point of view, we have proposed an image alignment method to make the computed multi-shot based signatures more reliable by taking and comparing the information from corresponding body parts on all images. We have also increased the discriminative power of RHSP patches by characterizing them using covariance descriptors, containing both color and texture information. We have added SIFT features to the final signatures, adding more texture information. We have also proposed a classification method for people visible side, based on our object tracking algorithm and the camera calibration information, allowing more precise comparison. We have finally proposed a method to use camera calibration information to reduce the number of candidate for re-identification and to weight some matching hypothesis using real world distance.

From processing time point of view, the baseline method [Farenzena 2010] is a real-time method for small sets of images per person (up to 8 images) and pseudo real-time method (up to 12 images). The improvements we have proposed does not increase greatly the processing time, and the final approach requires similar processing time. This is due to many reasons.

First, symmetry and asymmetry axes computation time is saved by the simple constant vertical division of the people bounding boxes. Second MSCR features are computed using some optimisations. In fact, after computing all MSCR features of the first image of a person, the MSCR of the following images are computed faster using better initialization parameters (centroid localisations, region areas and orientations, region colors) knowing the results of MSCR of the previous image and with the strong assump-

tion that the most important MSCR vary slightly from an image to the next one.

Even if covariance matrices are the most time consuming features, the simple computation and comparison still stay in real-time (or pseudo-real time) range. In people detection approach, the most consuming process step is the covariance mean computing, which is based on iterative gradient descent process. Here, this process is not required. Images alignment process is also fast thanks to the used downscale step and the use of calibration information for dimension range definition.

SIFT features detection and and their reliability measure computation is provided by our mono-camera tracking algorithm and do not have to be performed here. In the case of unavailability of our mono-camera tracking results, the SIFT features are detected on the last image of each person using the same detection and selection method as described in 5.2.2. This does not increase greatly the processing time. Finally, The same mono-camera tracking algorithm provides the real world trajectories which allows to perform the visible side classification and the spatio-temporal coherency filtering. The visible side classification consisting in simple regression function computation on subsets of the trajectory, the processing time is negligible.

The evaluation and comparison with state of the art results which validate our re-identification approach are provided in chapter 7, but some issues still exists with our method. The two main issues consists in the used features and also in the way they are weighted. The proposed adaptive feature weighing method, even it provides good results in our tests, is not the optimal method. Some importance scores are computed relatively (to the max distances of observed color, or to the max of observed entropy) and provides then a measure of discriminance in comparison with other people, but not a global measure on the richness in term of a given information type which may be more informative for better weighting. On the other hand, both color and textures may not be available in enough amount in some cases (dark and uniform clothes). This is a more general issue in state of the art and does not concerns only our approach.

7

EXPERIMENTAL RESULTS

This chapter presents a large evaluation of the proposed methods for people detection, mono-camera tracking and re-identification. It shows the performance improvements and the remaining limitations. Due to the unavailability of a whole evaluation framework (people detection, tracking, and re-identification) evaluation in the state of the art, the evaluations are performed for each processing task independently. We compare the results of each task with the state of the art results. The evaluation datasets are selected according to their popularity (the availability of results) and/or for the challenge they provide.

For each processing task, the corresponding evaluation metrics are firstly presented. Then, the state of the art datasets used for evaluation and benchmarking are presented. Finally, the evaluation results on each dataset are discussed, highlighting the improvement but also the limitations, explaining the most important reasons and proposing some ideas to solve the remaining issues for future work.

7.1 Efficient People Detector

7.1.1 Evaluation Metrics

There are two methodologies for people detector evaluation in the literature: the "***Per-Window (PW)***" performance evaluation and the "***Full Image***" performance evaluation. The "***Per-Window (PW)***" methodology is more dedicated to evaluate the performances of the classifiers only, by returning a decision concerning predefined candidate regions (selected by another process or delimited manually), while the "***Full Image***" methodology evaluates the whole detector, including the classifier (if it exists, because

some detectors are not classifer-based), the browsing image method, the scaling search, etc.

"***Full Image***" methodology provides a natural measure of error of an overall detection system, for this reason and because most state of the art approaches are evaluated using "***Full Image***" methodology, we use this methodology for our evaluations too. To decide if a region detected as a person is considered as a correct detection or not, the overlapping area between the detected bounding box ($\mathrm{BB}_{\mathrm{det}}$) and a ground truth one ($\mathrm{BB}_{\mathrm{gt}}$) has to be greater than a threshold. We use the PASCAL [Everingham 2009] measure, which states that the overlapping must exceed 50% as follow:

$$\frac{\mathrm{area}(\mathrm{BB}_{\mathrm{det}} \cap \mathrm{BB}_{\mathrm{gt}})}{\mathrm{area}(\mathrm{BB}_{\mathrm{det}} \cup \mathrm{BB}_{\mathrm{gt}})} > 0.5 \tag{7.1}$$

An efficient people detector has to detect the maximum number of appearing people (true positives or **TP**) in images while it has to make the minimum number of errors, i.e. the number of non-people detected as people (false positive or **FP**).

In literature, the first variable, i.e. true positive (TP) rate is generally replaced in evaluation by the false negative (FN) one, called also miss-detection rate. These two values are correlated since: TP rate $= 1.0 -$ FN rate

False positive and miss-detection rates are also correlated in a more general sense. In fact, the more a given people detector is "permissive" the lower the miss-detection rate is, but at the same time, the higher the false positive rate is too. In the other hand, the more this people detector is "restrictive", the lower the false positive rate is, but the higher the miss-detection rate is.

The evolution of miss-detection rate with respect to false positive one can be obtained by varying the permissivity of the evaluated people detector. This is done in several ways depending on the nature of the detector, even by varying the detection parameters (searching scales, browsing steps, width/height ratio of the detection window, etc...) of by varying the classifier parameters: for example, in the case of SVM classifiers, the permissivity variation can be obtained by shifting the separation hyperplan more in the positive class direction (increasing FN and decreasing FP rates) or the negatvive class one (increasing FP and decreasing FN rates. SVM hyperplan separation is explained in sec 2.1.3.2). For a boosted classifier, this permissivity variation can be obtained by varying the classifier threshold which is used for decision after summing the weak classifier responses. For a boosted cascade of classifiers, the threshold of each cascade level can be varied too, obtaining a variation in the permissivity of the detector. Another method consists in changing the number of considered cascade levels. By decreasing the number of cascade levels, the permissivity of the cascade is increased too (due to the rejection mechanism of this kind of classifiers).

Miss-detection and false-positive rates are given by the following equations:

$$\textbf{miss-detection rate} = \frac{N_{FN}}{N_{FN} + N_{TP}} \tag{7.2}$$

$$\textbf{false positive rate} = \frac{N_{FP}}{N_{FP} + N_{TN}} \tag{7.3}$$

where TP, TN, FP and FN are the numbers of true positive, true negative, false positive and false negative respectively. $N_{FN} + N_{TP}$ (the denominator of eq. (7.2)) corresponds to the number of annotated people in the ground truth while $N_{FP} + N_{TN}$ (the denominator of eq. (7.3)) corresponds to the number of tested negative regions.

To compare detectors we plot miss rate against false positives per image in log-log scale.

7.1.2 Dataset Presentation

The proposed people detector has been evaluated on four dataset: INRIA Person dataset, DaimlerChrysler dataset, Caltech Pedestrian dataset and CAVIAR dataset.

7.1.2.1 INRIA Person Dataset

The INRIA person dataset [Dalal 2005] consists of two subsets of color images: a training set containing 2416 person annotations and 1218 person-free images and a test dataset with 1132 persons and 453 person-free images. The pedestrian annotations were scaled into a fixed size 64×128 window, which includes a margin of 16 pixels around the pedestrians.

This dataset is quite challenging due to the various scenes, content, and persons appearence and poses (see some samples in figure 7.1).

Figure 7.1: Samples from INRIA dataset. The first row consists of some examples of annotated people. The second row consists of some croped negative samples with the same size as annotated people (64×128 pixels).

7.1.2.2 DaimlerChrysler Dataset

The DaimlerChrysler data set [Munder 2006] contains gray scale images with 4,000 pedestrian (24,000 with reflections and small shifts) and 25,000 non-pedestrian annotations. As opposed to the INRIA data set, nonpedestrian annotations were selected by a preprocessing step from the negative samples, which match a pedestrian shape template based on the average Chamfer distance score. Both annotations were scaled into a fixed size 18×36 window, and pedestrian annotations include a margin of 2 pixels around.

The data set is organized into three training and two test sets, each of them having 4,800 positive and 5,000 negative examples. The small size of the windows, combined with a carefully arranged negative set, makes detection on the Daimler- Chrysler data set extremely challenging. In addition, 3,600 person-free images with varying sizes between 360×288 and 640×480 are also supplied.

Figure 7.2: Samples from DaimlerChrysler dataset. The first row consists of some examples of annotated people. The second row consists of some croped negative samples with the same size as annotated people (18×36 pixels)

7.1.2.3 Caltech Pedestrian Dataset

The Caltech Pedestrian Dataset consists of approximately 10 hours of 640x480 30Hz video taken from a vehicle driving through regular traffic in an urban environment. About 250,000 frames (in 137 approximately minute long segments) with a total of 350,000 bounding boxes and 2300 unique pedestrians were annotated. The annotation includes temporal correspondence between bounding boxes and detailed occlusion labels.

This dataset is divided into a training and a test datasets. The training dataset consists of six training sets, each with 6-13 one-minute long sequence files, along with all annotation information. The testing data consists of five sets.

The sequences being acquired by a mobile camera, the annotated people are used as single images in our evaluation, due to the inability to use background subtraction

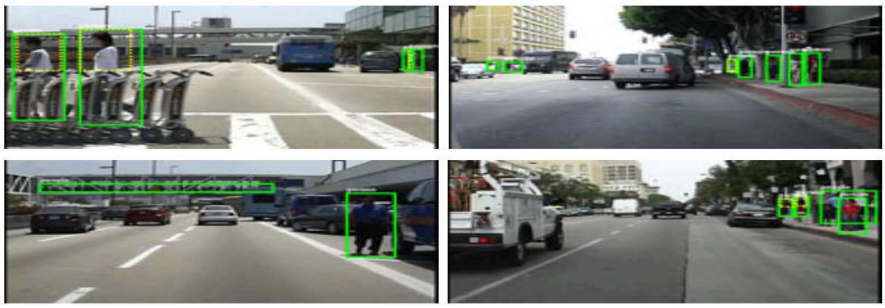


Figure 7.3: Samples from Caltech dataset.

information in the feature vector for covariance computation, in addition to the inability to use real world information (calibration) for candidate windows selection.

7.1.2.4 CAVIAR Dataset

The CAVIAR dataset consists of a set of 80 color video clips which were recorded acting out the different scenarios of interest. These include people walking alone, meeting with others, window shopping, fighting and passing out and last, but not least, leaving a package in a public place.

The CAVIAR dataset is divided into two sets of data. The first section contains 28 video clips which were filmed for the CAVIAR project with a wide angle camera lens in the entrance lobby of the INRIA Labs at Grenoble, France. The resolution is half-resolution PAL standard (384 x 288 pixels, 25 frames per second) and compressed using MPEG2. The second set of data also used a wide angle lens along and across the hallway in a shopping centre in Lisbon, Portugal. It contains 26 scenario. For each scenario, there are two time synchronised videos, one with the view across and the other along the hallway. The resolution is half-resolution PAL standard (384 x 288 pixels, 25 frames per second) and compressed using MPEG2.

Figure 7.4: CAVIAR dataset: (a) INRIA Grenoble sequences. (b,c) Lisbon shopping center sequences (Front and Corridor views respectively).

The sequences from INRIA Grenoble have been rotated by 90° (and the correspond-

ing ground truth data) in clockwise before processing, due to the roll angle of acquisition (see figure 7.4 (a)). Our people detector has been trained with images in which people have vertical pose or with small inclination (see figure 7.1; first row).

The interest of this dataset for people detection evaluation in comparison with the three previous ones is the availability of video sequences. This allows to uses background subtraction based features in the feature vector for covariance computation.

7.1.3 Evaluation Results

In the following evaluations, we have varied the miss-detection and false positive rates by removing iteratively the last level of our cascade of classifiers. The shorter the cascade is, the more permissive the detector is, i.e. it provides lower miss-detection rate while it increase false positive rate.

7.1.3.1 INRIA Dataset

This dataset being provided as single images acquired at different locations with different backgrounds, the background subtraction features ($\mathbf{G}$ and $\sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}$, see sec. 4.1.2) are not used in the covariance descriptors due to their unavailability. The camera calibration information is not available too, avoiding the use of real world information for candidate region selections. This increases the processing time of the detection process.

We have evaluated our method on INRIA People dataset and compare the obtained results first with the ones obtained with the baseline approaches ([Tuzel 2007] and [Yao 2008]) to quantify the improvements and second with the results of state of the art approaches.

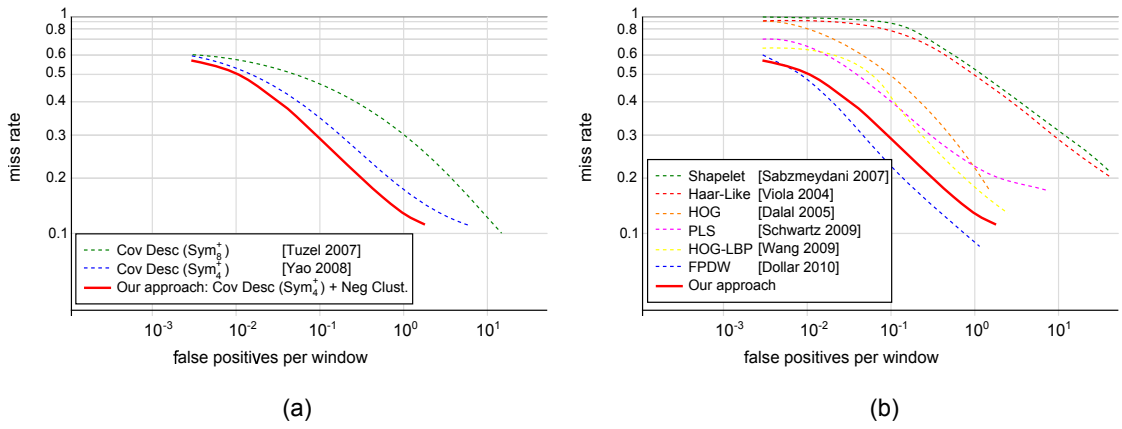


Figure 7.5: Results on INRIA dataset

Figure 7.5 (a) shows the performance comparison with [Tuzel 2007] and [Yao 2008] which have served as the basis of our work. Our method outperforms these two approaches. This is due to the more accurate strong classifiers at each cascade level. As mentioned in the explanation of our approach (see chapter 4), the random nature of candidate weak classifier selection in [Tuzel 2007] and [Yao 2008] does not allow to select the best possible weak classifier at each iteration. Our method limits the effect of the random selection by using groups of similar negatives (in terms of content with respect to the used region covariance descriptors) to train each cascade level. Thanks to this, the characterisation of positive data (people) is achieved more reliably.

Figure 7.5 (b) shows the performance comparison with state of the art approaches. The best performance on this dataset is achieved by our method and "The Fastest Pedestrian Detector in the West" (FPDW) approach [Dollar 2010], based on a trained classifier with Adaboost on gradients of both gray scale and color images. Our method provides the lowest false positive rate while it has the lower miss-detection rate in comparison with [Dollar 2010] ones. Our detector performs the detection with an average frame rate of 7 fps while [Dollar 2010] does it in 9 fps approximatively.

Note that by removing at least the two last levels of our cascade of classifiers, the evolution of the miss-detection rate against the false positive rate is better ensured by [Dollar 2010]. From a pure detection performance point of view, this last remark does not have any interest as long as the main objective is to obtain a detector with the lowest miss-detection and false positive rates, which is the case with our detector, but in practice, other parameters have to be taken into account to decide which detector is the best one for specific requirements. The most important parameter is the detection time. While [Dollar 2010] approach allows to vary progressively the permissivity of the classifier without impacting processing time (only by changing the classifier threshold), in our case, adding or removing some cascade levels change the processing time in a non-linear way with respect to the number of added/removed levels (due to the non-constant number of weak classifiers per cascade level). This means that depending on the processing time requirements of a given system and its tolerance to miss-detection and false positive rates, even our method or [Dollar 2010] one may be more adequate.

7.1.3.2 DaimlerChrysler Dataset

As for the INRIA People dataset, we have evaluated our method on DaimlerChrysler dataset and compared the obtained results first with the ones obtained with the baseline approaches [Tuzel 2007, Yao 2008] and second with the results of state of the art approaches.

Figure 7.6 (a) shows the performance comparison with [Tuzel 2007] and [Yao 2008].

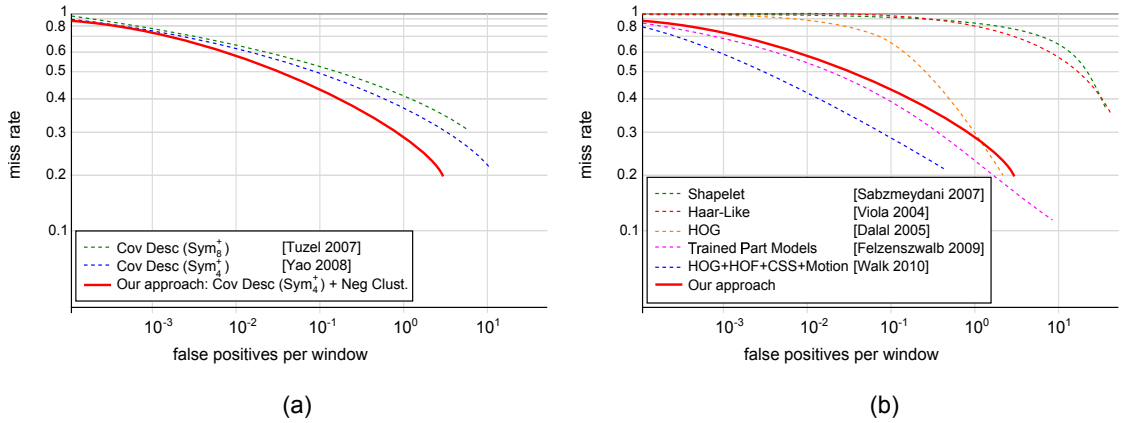


Figure 7.6: Results on DaimlerChrysler dataset

Our method outperforms these two approaches on this dataset for the same reason as for INRIA dataset, related to the more accurate strong classifiers at each cascade level.

Figure 7.6 (b) shows the performance comparison with state of the art approaches. The best performance on this dataset is achieved by [Walk 2010] approach, based on a combination of features (HOG, HOF, CSS, Optical flow) with a linear SVM classifier, and [**?**] approach, based on mixtures of multiscale deformable part models trained with a latent SVM.

Our method provides interesting results as long as they are not too far from those of the best approaches on this dataset. It outperforms the well-known [Dalal 2005] and [Viola 2004] approaches.

7.1.3.3 Caltech Pedestrian Dataset

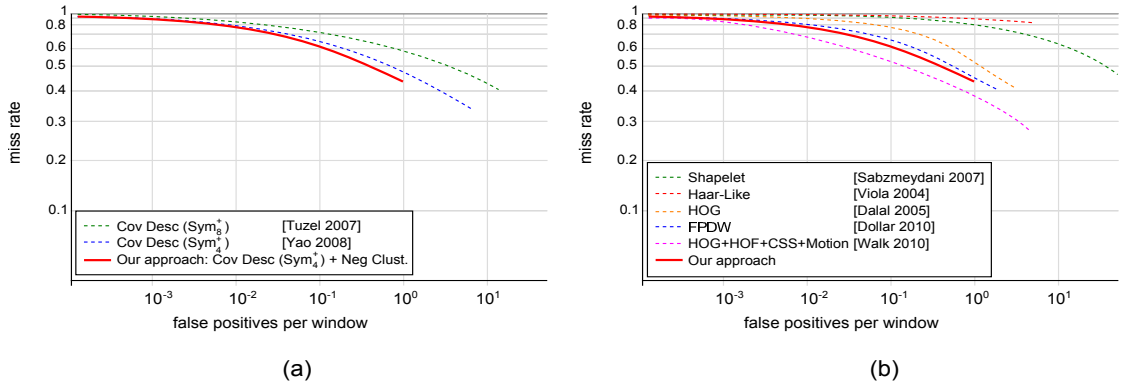


Figure 7.7: Results on Caltech dataset

We keep the same evaluation schema, by evaluating our method on Caltech Pedes-

trian dataset and by comparing our results first with the results obtained with [Tuzel 2007] and [Yao 2008] approaches on this dataset, and then with the results of state of the art approaches.

Figure 7.7 (a) shows the performance comparison with [Tuzel 2007] and [Yao 2008]. Our method outperforms these two approaches on this dataset.

Figure 7.7 (b) shows the performance comparison with state of the art approaches. The best performance on this dataset is achieved by [Walk 2010] approach too, as for DaimlerChrysler dataset.

Our method provides the second best results on this dataset. The most miss-detections in this dataset are due to partial occlusion. More than 50% of pedestrians are occluded on at least one frame. Our detector being a full body one, it is then more sensitive to occlusions with comparison with part-based detector.

7.1.3.4 CAVIAR Dataset

This dataset is provided as full video sequences, allowing the use of background subtraction features in region covariance descriptors. These sequences are acquired by static cameras, allowing their calibration and thereby the evaluation of the use of real world information for candidate region selection.

There is no available evaluation of state of the art people detection approaches on this dataset (initially, this dataset was used for tracking evaluation). For this reason, we have evaluated our detector and compared it with [Tuzel 2007] and [Yao 2008] approaches only.

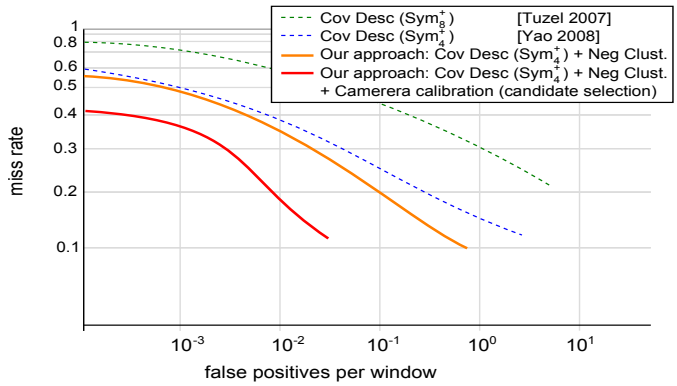


Figure 7.8: Results on CAVIAR dataset

Figure 7.8 shows the results of [Tuzel 2007] and [Yao 2008] approaches, as well as the results of two versions of our detector: the first version consists of our detector without using camera calibration information for candidate region selection while the

second one uses the camera calibration information.

We can observe that even without using the camera calibration information, our detector outperforms the [Tuzel 2007] and [Yao 2008] ones. The reason is due to the negative sample clustering to reduce the effect of random selection of candidate weak classifiers, as explained above.

The most interesting remark concerns the contribution of camera calibration information. In fact, using this information allows to strongly reduce the number of tested regions during detection. The first effect is an important processing time reduction (around 60% in average in this dataset, but may vary from a camera to another one, depending on the height of the camera and the view angle). The second effect, which is visible on the curve is an important improvement of detection performances (less miss-detection with respect to false positives). This last effect is explained by the fact that less negative candidate regions are tested. During the detection process without using camera calibration information, several scales of testing windows are used. For each scale, several width/height ratios are used too (around a mean ratio of 1/3). Camera calibration information allows to reduce not only the number of localisation to test (according to the ground floor) but also the range of scales and width/height ratio for each location. This leads to a lower rate of false positives.

7.1.3.5 Dataset Dependency of the Detector

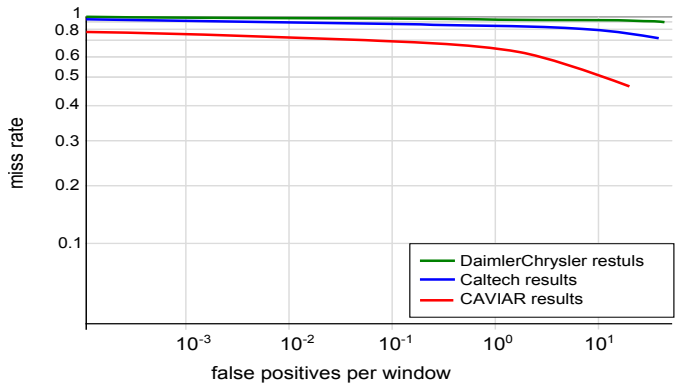


Figure 7.9: Results of the trained detector on INRIA dataset, applied on DaimlerChrysler, Caltech and CAVIAR datasets.

We have conducted a last experimentation to test how a trained detector on a given dataset is efficient on other datasets. To do this experimentation, we have selected the trained detector on INRIA dataset, due to the best results it provides in comparison with our other trained detectors on other datasets (except CAVIAR dataset, but we decide to

do not select it because it is background subtraction features based).

We have applied this detector on the test images of each other dataset (Daimler-Chrysler, Caltech and CAVIAR). The evaluation results are shown in figure 7.9. We can observe that the detection performances are very bad. This demonstrate the dependency of a people detector to the dataset which has served for its training.

We are convinced that this is not an overfitting issue since the application of this detector (trained on INRIA dataset) on several images which are found on internet and not in the INRIA dataset (see figure 7.10) still provides good results. We believe that some specific characteristics of each dataset are not managed well during training stage (in all state of the art approaches) and are not integrated as full part features.

Figure 7.10: Examples of detection results using our people detector trained on INRIA dataset, applied on several images that do not belong to INRIA dataset.

7.2 Robust People Tracking Using Particle Filter

7.2.1 Evaluation Metrics

Object tracking performances can be evaluated under several criteria, according to all possible situations which can occur during the tracking. Most generally, an "efficient" tracking algorithm must track **all** objects of interest in the scene, it must start tracking them as soon as possible after they appear in the scene, it must maintain the tracking of these objects are in the scene and does not stop the tracking until they leave the scene. It must also assign a unique ID to each tracked object all the time this object is in the scene (not necessary visible at all the time: occlusion management).

Using ground truth data, several metrics are proposed in state of the art to quantify how a given tracking algorithm complies with all or a part of these conditions.

In our evaluation, we use two types of metrics to compare our results on state of the art results:

7.2.1.1 ETISEO metrics

ETISEO dataset, as it is presented in the next section, is rich in terms of environments and external conditions, so it is an interesting dataset to evaluate our tracking algorithm. In order to be able to compare our tracker performances with the other ones on the ETISEO videos and on Caretaker dataset (presented in the next section), we use the tracking evaluation metrics defined in the ETISEO project [Nghiem 2007]. These metrics allows to quantify precisely how the trackers comply with all the previously mentioned conditions.

In the following paragraphs, "reference object" refers to an annotated object in ground truth (object to track), and "tracked object" refers to an object delimited and tracked by the evaluated tracker. The match between a reference object and a tracked object is done with respect to their bounding boxes.

ETISEO evaluation metrics for object tracking consists of three metrics:

- **Tracking time** metric, denoted $\mathbf{M_1}$, measures the percentage of time during which objects of interest (reference objects) are really tracked. It is given by:

$$M_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i}{F_i} \tag{7.4}$$

where N is the number of reference objects; T_i is the number of frames in which the reference object i is tracked (by the considered tracking algorithm); F_i is the number of frames for which the reference object i is annotated.

- **Object ID persistence** metric, denoted $\mathbf{M_2}$, is used to evaluate the ID persistence. It computes over the time how many tracked objects are associated to one reference object as follows:

$$M_2 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{Frag_i} \tag{7.5}$$

where N is the number of reference objects, $Frag_i$ is the number of different tracked objects (provided by the evaluated tracker) which corresponds to the reference object i.

- **Object ID confusion** metric, denoted $\mathbf{M_3}$, computes the number of reference object IDs per detected object as follows:

$$M_3 = \frac{1}{D} \sum_{i=1}^{D} \frac{1}{Conf_i} \tag{7.6}$$

where D is the number of tracked objects (provided by the evaluated tracker) matching with ground truth data, $Conf_i$ is the number of different reference objects which corresponds to the tracked object i.

Each of these metrics helps to evaluate a given aspect of the tracking algorithm efficiency. Their values are in the interval [0, 1]. The higher the metric value is, the better the tracking algorithm performance gets.

7.2.1.2 MT, PT and ML metrics

Wu et al. have defined in [Wu 2007] five other metrics to quantify differently how a given tracking algorithm complies with the criteria mentioned at the beginning of this section. Three of these metrics are mainly used in several state of the art publications [Xing 2009, Huang 2009, Huang 2008, Li 2009, Kuo 2010, Chau 2011]. Let Nb_{GT_Traj} be the number of trajectories in the ground-truth of the considered video sequence. These three metrics are defined as follows :

- **Mostly tracked trajectories (MT) :** it correspond to the number of trajectory which are tracked correctly for more than 80% divided by Nb_{GT_Traj}.

- **Partially tracked trajectories (PT) :** it correspond to the number of trajectory which are tracked between, 20% and 80% divided by Nb_{GT_Traj}.
- **Mostly lost trajectories (ML) :** it correspond to the number of trajectory which are tracked less than 20% divided by Nb_{GT_Traj}.

For each GT object, the trajectory of the considered tracker which overlaps the one of GT object for most of the time is considered as the correct one.

7.2.2 Dataset Presentation

We have conducted the experimentation of our object tracking algorithm on four datasets: PETS 2001, ETISEO, CAVIAR and Caretaker.

7.2.2.1 PETS 2001 Dataset

The PETS 2001 dataset consists of five separate sets of training and test sequences, containing outdoor people and vehicles. All the datasets are multi-view (2 cameras) and provide significant lighting variation, occlusion, scene activity and use of multi-view data. Only the 3 first sets are used for our evaluation. The two last sets are out of scope for our study. In fact, the 4th dataset contains sequence provided by fisheye camera while the 5th one contains sequences provided by moving camera, fixed on a car and acquiring video on the road.

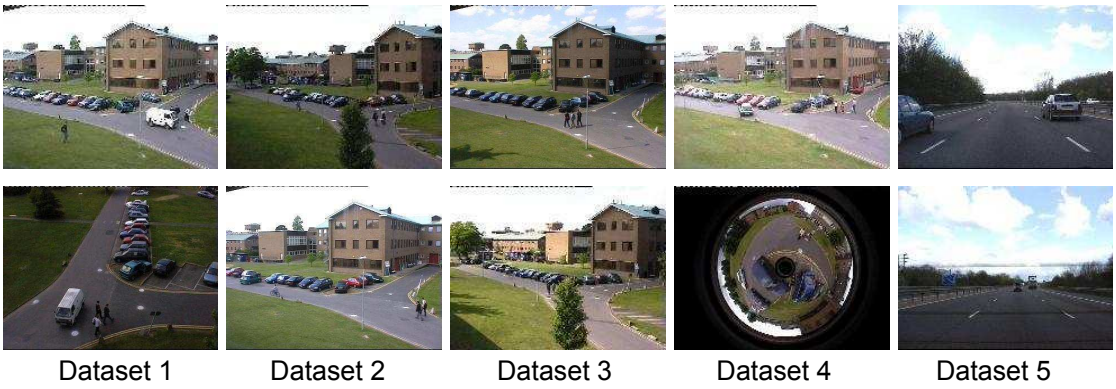


Figure 7.11: PETS 2001 dataset.

7.2.2.2 ETISEO Dataset

The ETISEO videos are provided by the ETISEO project. This project seeks to work out a new structure contributing to an increase in the evaluation of video scene under-

standing ; with the active participation of industrial and many research laboratories, such as French, European and International research centers.

Project ETISEO focuses on the treatment and interpretation of videos involving pedestrians and (or) vehicles. The dataset contains 86 video clips. These sequences constitute a representative panel of different video surveillance areas. They merge indoor and outdoor scenes, corridors, streets, building entries, subway station, etc. They also mix different types of sensors and complexity levels. This makes this dataset relatively challenging and allows to evaluate the generecity of the proposed tracking algorithms.

The ETISEO project provides a set of tracker evaluation results. We test the proposed mono-camera tracking algorithm on this dataset and we compare our results with these available results and with more recent trackers results which are evaluated on this dataset.

Figure 7.12: Samples from ETISEO dataset

7.2.2.3 CAVIAR Dataset

The Caviar dataset is presented in sec. 7.1.2.4. Both INRIA Grenoble and Shopping center of Lisbon are used for the evaluation of our object tracking algorithm.

The INRIA Grenoble sequences are used as they are provided, without any rotation to obtain vertical people like for people detection evaluation. This is a deliberate choice, to show the efficiency of our approach and the used particle filter for object tracking, independently of flat world constraints, as long as the calibration is correctly done. In the ambiguous cases, where the classification by the real dimensions is not reliable (see explaination in sec 5.1.2 and figure 5.8), we apply a rotation of 90° to the ambiguous object (unclassified bounding box) and we apply our people detector as explained in sec. 5.1.2. This does not happen frequently in this dataset, so the real-time processing is not affected.

7.2.2.4 Caretaker Dataset

The Caretaker project focuses on the extraction of a structured knowledge from large multimedia collections recorded over networks of cameras and microphones deployed in subways. This dataset is highly challenging due to many factors: the scenes are crowded most of the time, the light reflection on smooth floor deteriorates the background subtraction providing imprecise people detection and delimitation, the poor video quality (highly compressed data), and the numerous static and dynamic occlusions. The frame rate of the videos (5 frames per second) also makes this dataset challenging for our proposed approach (due to the prediction-update nature of our approach).

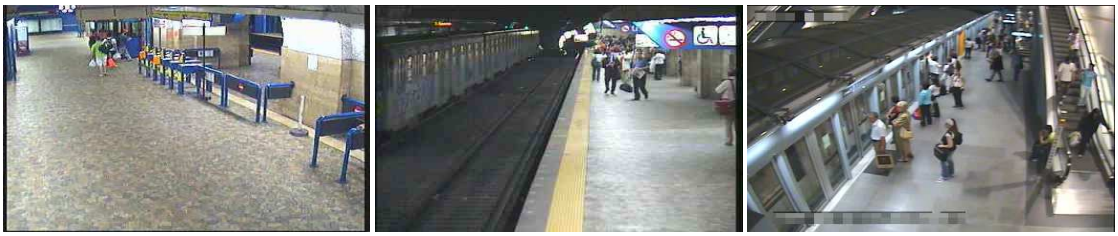


Figure 7.13: Samples from Caretaker dataset.

7.2.3 Evaluation Results

7.2.3.1 Comparative Evaluation on ETISEO Dataset Using ETISEO Metrics

We have evaluated our mono-camera tracking algorithm on ETISEO dataset. This dataset has been acquired in several indoor/outdoor places and it contains various environements. It has been proposed for ETISEO project and has been used by several teams to evaluate their detection and tracking algorithms. Unfortunately, the names of these teams are not available and they are identified by numbers in the ETISEO project to preserve their anonymity. Details about their tracking approaches are also missing. Available ETISEO project data provides only the final results.

We have compared our tracking algorithm with those of 7 teams from the project and with two tracking algorithms proposed in [Chau 2011] on two sequences of the dataset. The first sequence, denoted ETI-VS1-BE-18-C4 shows a building entrance. This sequence provides low difficulty level. The second sequence denoted ETI-VS1-MO-7-C1 shows an underground station. It is more challenging since it contains occlusions and is acquired with low contrast and bad illumination.

The results of this comparison is shown in figure 7.14.

Our tracking algorithm outperforms all other evaluated algorithms (or is equal to the best ones in some cases) except for the **Object ID persistence** metric ($\mathbf{M_2}$) in the

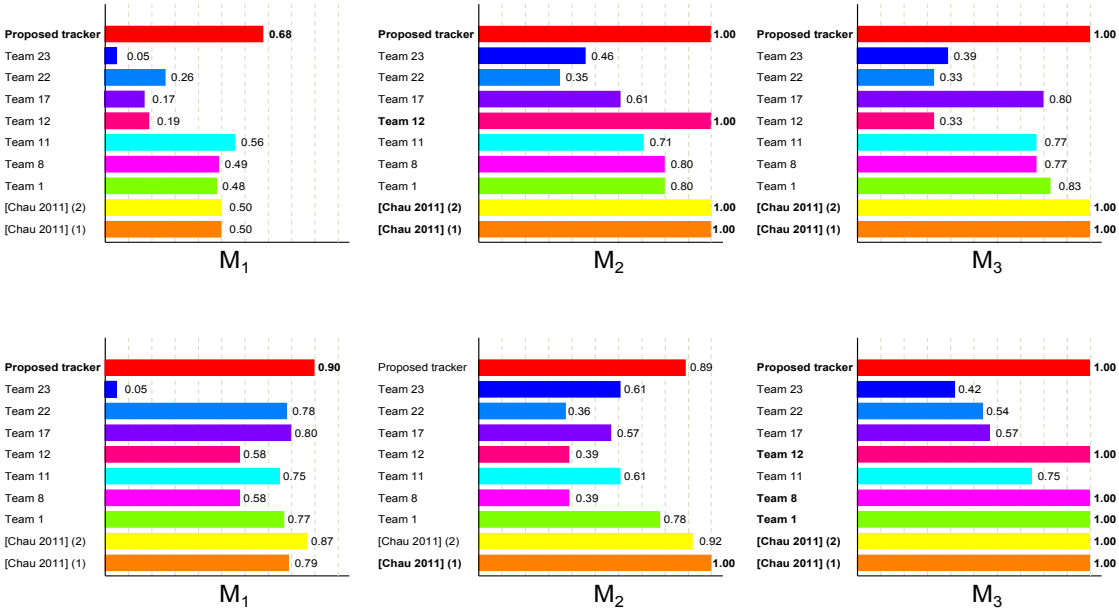


Figure 7.14: Detailed results and comparison on two sequences from ETISEO dataset. First row: ETI-VS1-BE-18-C4 sequence results. Second row: ETI-VS1-MO-7-C1 sequence results. Bold text corresponds to the approach which provides the best results per metric. [Chau 2011](1) refers to an off-line learning based tracker; [Chau 2011](1) is and estimator-based tracker.

underground station sequence. For this metric, [Chau 2011](1) tracking algorithm performs better. This is due to the crowd environment in this sequence. [Chau 2011](1) tracking algorithm is based on an off-line learning to extract the best features and parameters. In opposite, our algorithm does not use any off-line learning and thereby, in on-line processing, it does not have enough time to learn correct reliability measures for each tracked SIFT features in one hand, and to learn correct real world dimensions and velocity to manage occlusions. More ID switches happen due to the unreliability of these measures.

A global evaluation result on the whole dataset is shown in figure 7.15.

7.2.3.2 Evaluation on ETISEO, PETS 2001, CAVIAR and Caretaker Sequences Using ETISEO Metrics

We have also evaluated our mono-camera tracking algorithm on PETS 2001, CAVIAR and Caretaker datasets. The global results are shown in figure 7.15.

We can see that the results on PETS 2001 and CAVIAR are quite close to each other and close to the global results on ETISEO dataset, except for $\mathbf{M_2}$ on ETISEO (see figure 7.15). The close results are explained by the fact that even if the environments and

sequences of ETISEO, CAVIAR, and PETS 2001 are different in terms of color rendering and illumination conditions, they have similar/close resolution, people size on images and the most important common point is that they do not contain crowded scenes (except for the underground station sequences from ETISEO).

Our tracking algorithm does not use color information for tracking, and uses it for occlusion management in a simple way (dominant color, see sec. 5.4.2). The SIFT features have proven their robustness against illumination changes, so the context of these three datasets is quite similar with respect to the nature of our tracker (SIFT features + particle filtering).

The low value of $\mathbf{M_1}$ on ETISEO dataset (see figure 7.15) is mainly due to the underground station sequences which decrease the global mean values of metrics on the whole datasets. In these crowded sequences, the large amount of occlusions, especially at the begining of sequences delays the tracking start, decreasing the final tracking time.

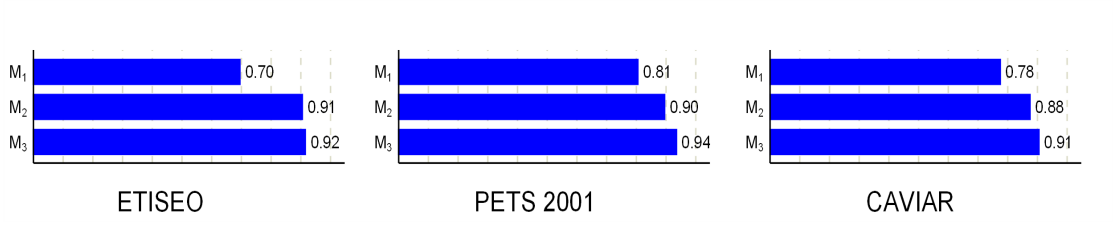


Figure 7.15: Global results of the proposed approach on ETISEO, PETS 2001 and CAVIAR datasets.

The same reason (crowd scene) explains the low score of $\mathbf{M_1}$ and $\mathbf{M_2}$ for Caretaker sequence (see figure 7.16), concerning the initial tracking algorithm (red bars, without offline learning). In fact, the crowded scenes at some moments of the sequence does not provide enough time to our tracking algorithm to learn (on-line) the correct reliability measures of each SIFT feature, to allow dynamic models of SIFT features to converge (for particle filtering) and to learn the real dimensions and velocity of tracked people. This issue concerns the new appearing people in the scene directly in crowded situations. People who appear relatively separated from crow for enough time are well managed by our tracker event if they interact with the crowd later.

Even if our presented tracking system does not include any offline learning step to be deployed, we have conducted a special experiments to highlight the fact that our tracker performs better if complex situations (crowd/oclusions) does not occure immediately when a new object of interest appear in the scene, i.e. if the tracker has enough time to "learn" SIFT features reliability measures, dynamic models and object real world dimensions and velocity in online mode. To simulate this "enough time" requirement, we

have modified slightly our tracking algorithm to allow it to load tracking ground truth data from xml file when they are available and to use them exclusively as the tracking result at each frame, changing all the variables and parameters values accordingly (use ground truth objects as delemitation instead of background subtraction/people detector results, correction/removing of SIFT features if they do not cooresponding to the tracking state, correct real world dimensions and velocity after ground truth data projection using camera calibration information, etc.). When the ground truth data of a given object at a given moment are no longer available, the tracker resumes it process for this object normally as it was desinged initially.

After that, we have created a partial ground truth data by taking the original dataset ground truth and by keeping only the first 10 to 15 frames data for each object of interest. This correspond to 2 to 3 seconds (this dataset is acquired at 5 frame per second).

Finally, we load this partial ground truth in our modified tracking algorithm and we start the processing (object tracking on this sequence). For the evaluation, we did not count the annotated frames of the partial ground truth in the metrics computations (it is not fair to count them as long as the output of these frames is exactly the ground truth data for each involved object), we start metric computation for each object at the first frame the tracker resumes the tracking with autonomy. The results of this evaluation are presented in figure 7.16 at the upper bar (pink bar). We can observe that if our tracking algorithm has enought time (10 to 15 frame in our experiments) to learn and stabilize the several parameters involved in the process, which occures autonomously and online (without ground truth data) if objects of interest appears and evolve without occlusion in few frames, the results of tracking may be much better. Unfortunately, the offline learning is not a part of our proposed tracking algorithm, due to the hard application in real deployed video-surveillance systems.

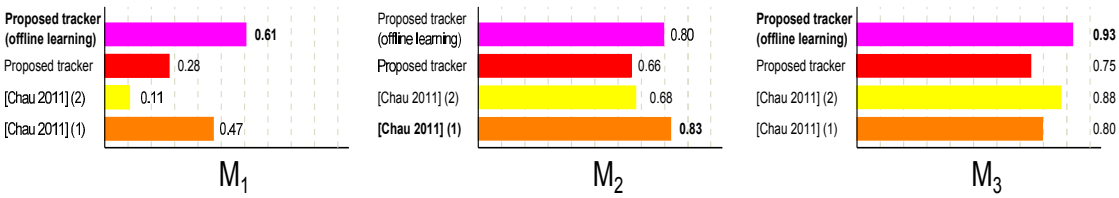


Figure 7.16: Comparative results on Caretaker sequence. [Chau 2011](1) refers to an off-line learning based tracker; [Chau 2011](2) is and estimator-based tracker.

7.2.3.3 Comparative Evaluation on CAVIAR Dataset Using MT, PT and ML Metrics

We have evaluated our tracking algorithm on CAVIAR dataset using [Wu 2007] metrics to be able to compare our results with state of the art trackers. The results are shown

in figure 7.17.

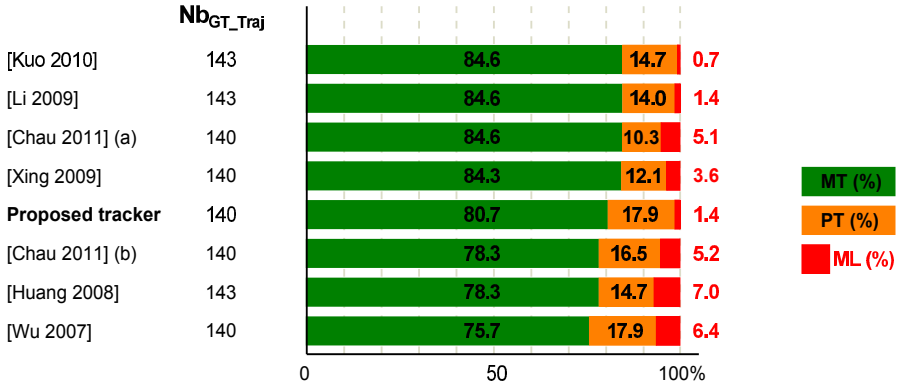


Figure 7.17: Comparative results on CAVIAR dataset. [Chau 2011](a) refers to the tracker with the proposed controller; [Chau 2011](b) is the tracker without the proposed controller.

We can observe that our tracking algorithm provides good results, but it is outperformed by some state of the art tracking algorithms. The partially tracked and mostly lost people are due to the image noise and low resolution. In fact, people tracking performances depend on people detection, and in this conditions, background subtraction and people detection results are altered, impacting the tracking start of some people, especially those who come from the far end of the corridor, and then decreasing the time these people are correctly tracked.

On the other hand, the best results are provided by [Kuo 2010], [Li 2009] and [Chau 2011] (a) trackers. The higher performances of [Li 2009] approach are due to the off-line learning step (HybridBoost) performed to select the best parameters and/or the most discriminative features. This adapts the tracker for this specific sequences. Our aim being to provide a tracker which is as generic as possible and which may be deployed with a minimum parameter tuning stage by operators, this explains the lower performance of our tracking algorithm, even if they are not bad.

[Chau 2011] better performances are due also to an off-line learning, but not only. [Chau 2011] propose an hybrid-based (on line/off-line) controller to select the best features and parameters and to refine the parameters on-line according to the tracking performances and scene context changes. The on-line part of this controller may be an interesting way to investigate to improve our tracker performances without decreasing its genericity.

7.3 Fast People Re-Identification

Most of datasets for benchmarking consists of extracted (cropped) images of people of interest. Some datasets provides only one image per person and per camera while the others are provided without real world information (real world coordinates and velocity), preventing the use of some of our improvements, especially the visible side classification and the spatio-temporal coherency filtering.

However, the other improvements are evaluated and compared to the initial approach of [Farenzena 2010] and to the other state of the art approaches.

7.3.1 Evaluation Metrics

The re-identification system performance is strongly dependent on the number of considered candidate people for any re-identification query. The larger the candidate number is, the higher the probability to have similar people than the query person (in terms of appearance) is, and thereby, the lower the probability to find the corresponding person among the candidate ones is. It is then necessary to associate a performance measure to the number of considered people.

For assisted people tracking under camera network, or for an off-line system of people retrieval on stored videos, re-identification systems may significantly decrease the number of possible matching per query person, by filtering the less-similar people and by proposing the most “n” likely ones for the real matching (n depends on the size number of considered people in all the available video sequences and the expected performances). This helps operators to speed up their tracking/search of people of interest.

For these reasons, the most used metric for re-identification system evaluation is the Cumulative Matching Characteristic (CMC) curve.

7.3.1.1 Cumulative Matching Characteristic (CMC) Curve

For re-identification performance evaluation, we use a metric known as the cumulative matching characteristic (CMC) curve (see figures 7.18). The CMC curve represents the expectation of finding the correct match in the top n matches. The best case is given by a re-identification rate of 100% at the first rank (see figure 7.18 (b)). The higher the re-identification rate is at the lower ranks, the more performant the re-identification algorithm is.

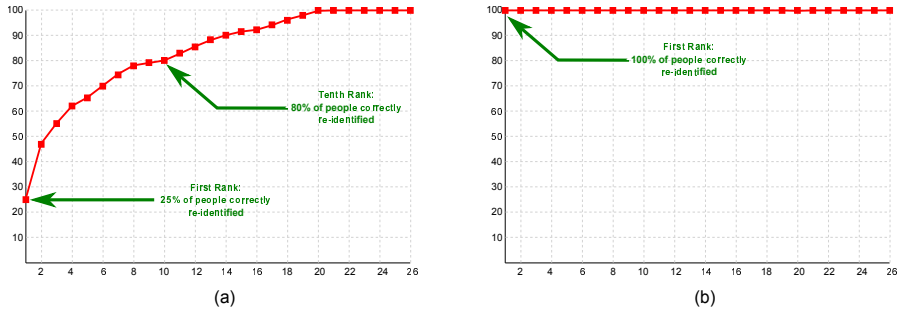


Figure 7.18: Cumulative matching characteristic (CMC) curve illustration. (a) Usual form of a CMC curve with with 25% of correct re-identification at the first rank and 80% of correct re-identification at the tenth rank; (b) The best (ideal) case with 100% of correct re-identification at the first rank.

7.3.1.2 Normalized Area Under Curve (nAUC)

The cumulative matching characteristic (CMC) curve provides a detailed performances per re-identification rank. To be able to evaluate and compare several re-identification algorithms, [Bazzani 2012] has introduced the "Normalized Area Under the Curve" (nAUC) value which provides re-identification performance measure as a scalar. It represents the area under the CMC curve expressed in % (see figure 7.19). The best case of 100% of re-identification at the first rank corresponds to $\mathrm{nAUC} = 100\%$ (see figure 7.19 (b)). The higher the re-identification rate is at the lower ranks, the larger the area under the CMC curve is, and thereby, the higher the nAUC value is, corresponding to the better re-identfication performances.

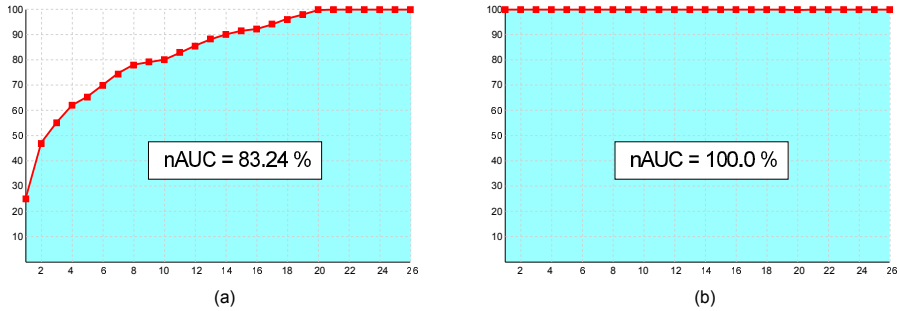


Figure 7.19: Corresponding nAUC value to the previous CMC curves (figure 7.18). (a) nAUC = 83.24% for the first case; (b) nAUC = 100% for the second case.

The nAUC values are not always provided by authors for their appraoches. For this reason, we will indicate the nAUC values only in two cases: If they are provided by the authors or if the published CMC curves reach 100% of re-identifcation in the provided figures (for large datasets, only a partial range of first ranks is displayed in published

CMC curves, and in these ranges, curves do not reach 100% of re-identification performances). For the second case, we provide a nAUC value thanks to a geometry software which allows us to estimate the area of the polygon delimited by the provided curve and its straight line extrapolation (knowing the size of the dataset and the fact that once a curve reach 100% it will be constant at this value for the remaining ranks).

7.3.2 Dataset Presentation

7.3.2.1 VIPeR Dataset

The VIPeR contains two views of 632 pedestrians. Each pair is made up of images of the same pedestrian taken from different cameras, under different viewpoint, pose and light conditions. All images are normalized to 128×48 pixels. The dataset contains pairs which viewpoint angle changes from 45 up to 180 degrees. Each pair is randomly split into two sets: CAM A and CAM B. The video was compressed before pairs of image extraction, so many compression artefacts are present on images.

Considering images from CAM B as the gallery set, and images from CAM A as the probe set, each image of the probe set is matched with the images of the gallery. This provides a ranking for every image in the gallery with respect to the probe.

This dataset is the most challenging dataset currently available for the single-shot human re-identification, but unfortunately, it does not correspond to a video surveillance case for two main reasons: first, only one image per camera is available in this dataset while many images per person and per camera may be provided by a surveillance system, and the acquisition position of the cameras is quite low, providing frontal view while in surveillance context, cameras are frequently higher and provide angled view acquisition.

(Note: the random assignment of images to one set or to the other one makes the colorimetric calibration useless (erroneous) here.)

Figure 7.20: Samples from VIPeR dataset

7.3.2.2 i-Lids Dataset

The Imagery Library for Intelligent Detection Systems (i-LIDS) is the U.K. government's benchmark dataset for video analytics (VA) systems. It has been collected by the Centre for Applied Science and Technology (CAST) in partnership with the Centre for the Protection of National Infrastructure (CPNI).

i-LIDS comprises a library of CCTV video footage based around "scenarios" central to the government's requirements. The footage accurately represents real operating conditions and potential threats.

i-LIDS provides two types of scenario datasets: the **event detection scenarios** which consists of sterile zone, parked vehicle, abandoned baggage, doorway surveillance, and new technology scenarios, and the **tracking scenario** which contains the multiple camera tracking scenario (MCTS) which is our scenario of interest, and which contains approximately 50 hours of footage provided by 5 cameras deployed in an airport, providing very challenging situations: many occlusions and large variation in appearance from a camera to another one, due to people wearing backpacks, carrying luggages or pushing carts.

Four subsets of images have been extracted for re-identification evaluation:

- **i-LIDS-119**
  This evaluation dataset contains 476 images with 119 individuals (each individual is represented by an average number of 4 images). These images have been extracted automatically by [Zheng 2009] from the sequences provided by two different cameras.

Figure 7.21: Samples from i-Lids-119 dataset

- **i-LIDS-MA (Manually Annotated)**
  Due to the low number of images per person in **i-LIDS-119** dataset (4 images in

average), which is not sufficient to exploit the advantages of using multiple images in generating human signature, Bak ([Bak 2011]) has extracted two other subsets of images for re-identification evaluation: i-LIDS-MA and i-LIDS-AA datasets.

i-LIDS-MA dataset contains 40 individuals extracted **manually** from two cameras. For each individual 46 frames have been annotated manually for both cameras. Therefore this dataset contains $40 \times 2 \times 46 = 3680$ annotated images.

This dataset provides images where in most of which, people are well delimited and centred in images.

Figure 7.22: Samples from i-Lids-MA dataset

- **i-LIDS-AA (Automatically Annotated)**
  The manually annotated dataset (i-LIDS-MA) does not reflect real video surveillance scenario where humans are detected and tracked automatically. Consequently, Bak et al. ([Bak 2011]) have extracted a new subset of people images automatically by applying HOG-based human detector and tracker to obtain multiple images of individuals seen from both cameras. In this case, detection and tracking results are noisy which makes the dataset more challenging. This dataset contains 100 individuals. For each individual, a different number of frames is automatically extracted, depending on tracking difficulties. In total, the dataset contains 10754 images.

- **i-LIDS-AA-RP (Automatically Annotated with Rotating People)**
  This dataset contains 30 individuals extracted automatically using our people detector (chapter 4) and mono-camera tracking (chapter 5) algorithms, providing multiple images for each person. The aim of this dataset, in comparison with i-LIDS-AA one, is the fact that only people who change significantly their walking direction while they are observed by the same camera are selected. This is a more

Figure 7.23: Samples from i-Lids-AA dataset

general case in comparison with those used in state of the art for re-identification evaluation. It allows us to evaluate the visible side classification contribution.

7.3.2.3 ETHZ Dataset

This dataset is captured from a moving camera, and it has been used originally for pedestrian detection [Ess 2007]. Schwartz and Davis in [Schwartz 2009] extract a set of samples for each different person in the videos, and use the resulting set of images to test their Partial Least Squares Analysis method for re-identification. The moving camera setup provides a range of variations in people appearance. Variation in pose is relatively small, though, in comparison with the other two datasets. The most challenging aspects of ETHZ are illumination changes and occlusions. Using the same camera to acquire all images deprives this dataset of a main challenge which is largely met in video surveillance context: rendering variation due to sensor difference.

All images are normalized to 64×32 pixels. The dataset is structured as follows:

- SEQ. #1 contains 83 pedestrians, for a total of 4.857 images.
- SEQ. #2 contains 35 pedestrians, for a total of 1.936 images.
- SEQ. #3 contains 28 pedestrians, for a total of 1.762 images.

7.3.2.4 CAVIAR4REID Dataset

The CAVIAR dataset is presented in sec. 7.1.2.4. For re-identification evaluation, only Lisbon shopping center sequences are used due to the availability of synchronized video sequences provided by two cameras with overlapping field of view (INRIA Grenoble sequences are provided by a unique camera).

Figure 7.24: Samples from ETHZ dataset

Cheng et al. ([Cheng 2011]) have created a new re-identification dataset called CAVIAR4REID by extracting a set of 50 individuals observed by both cameras, using the ground truth data provided by CAVIAR project. Each individual is represented by a set of 10 images selected by maximizing the variance with respect to resolution changes, light conditions, occlusions, and pose changes.

This dataset is very challenging due to the perpendicular view axes of the two cameras. People are mainly observed in front/back side in one camera while they are mainly observed from profile side in the other camera.

7.3.3 Evaluation Results

In the following paragraphs, we detail the evaluation of our approach on several dataset and compare our results with state of the art algorithms. We provide the results in two forms: a CMC curve displaying the results on a range of ranks which is common to all the state of the art approaches with which we compare our results, and a table containing numerical values of correct re-identification rates on the first and the 5th ranks in addition to the nAUC values when they are available/estimable. "–" means that the nAUC is not available/estimable.

7.3.3.1 VIPeR Dataset

We have tested our approach on VIPeR dataset and compared our results with state of the art ones. The results are shown in figure 7.25 and Table 7.1. Our approach outperforms the initial approach proposed by [Farenzena 2010] thanks to the use of SIFT features and covariance descriptors for Recurrent High-Structured Patches (RHSP) characterization. Unfortunately, our approach is outperformed by [Dikmen 2011] and [Cheng 2011] approaches. VIPeR dataset consists of a set of single-shot images per

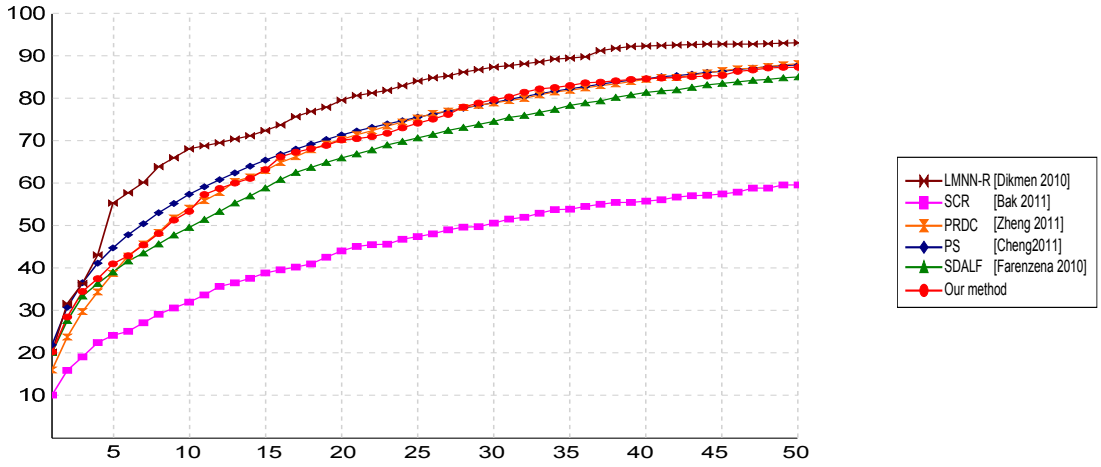


Figure 7.25: CMC curves obtained on VIPeR dataset.

person and per camera. This reduces strongly the effectiveness of our approach even it is designed for both single-shot and multiple-shot cases. Image alignment contribution cannot be applied in this case since it requires multiple images per person, reducing the performances of our approach. In addition to that, people images are acquired under viewpoint angle changes from 45 up to 180 degrees. With single image per person and without any real world information (the dataset is provided as cropped images without any additional information), the use of SIFT features and covariance descriptors for RHSP characterization by default weights (see sec. 6.3.7), may alter the final result by providing bad scores for people observed from different sides. The best performance on this dataset is achieved by LMNN-R [Dikmen 2010] and PS [Cheng 2011] approaches. Both approaches use color histograms as a feature representation; LMNN-R is based on metric learning while PS use body-part segmentation method.

Approach	**1st rank (%)**	**5th rank (%)**	**nAUC (%)**
PS [Cheng 2011]	21.93	44.73	93.60
Our approach	20.25	40.98	91.45
LMNN-R [Dikmen 2011]	20.22	55.27	–
SDALF [Farenzena 2010]	20.12	39.06	89.90
PRDC [Zheng 2011]	16.03	38.72	–
SCR [Bak 2011]	10.13	24.17	80.75

Table 7.1: Detailed results on VIPeR dataset.

7.3.3.2 i-Lids Dataset

- **i-LIDS-119**

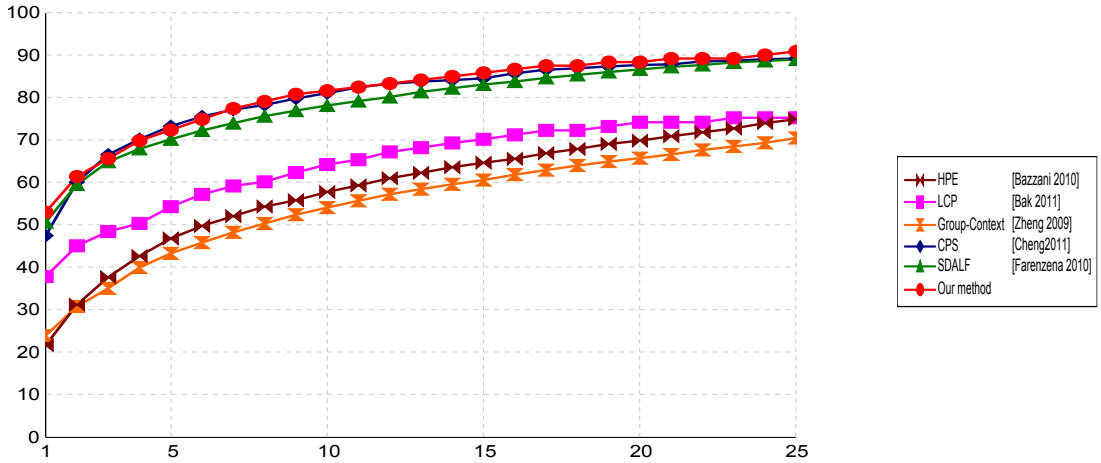


Figure 7.26: CMC curves obtained on iLids-119 dataset.

This dataset is very challenging since the number of images per individual is very low (4 images in average), in addition to the many partial occlusions it contains. It has been used to evaluate single-shot approaches as long as multiple-shot approaches. Most of state of the art authors have proposed approaches for both cases, and the evaluation of their approaches has shown that the multiple-shot approaches provide better results. For this reason, we will focus the comparison on the multiple-shot case since it corresponds to the best results that each author has obtained.

The results of our evaluation on this dataset, and the comparison to the state of the art approaches are shown in figure 7.26 and Table 7.2. We can observe that the best results are provided by our method and [Cheng 2011] (CPS: Custom Pictural Structure) one. Our method outperforms the initial approach of [Farenzena 2010] tanks to the several improvements we propose, especially the use of SIFT features, the covariance descriptors for RHSP characterisation, and a better feature weighting instead of those which are fixed and used by [Farenzena 2010] and which have been learnt from VIPeR dataset, demonstrating that some parameters are dataset-dependent.

Another contribution which may justify the better results obtained by our approach is the image alignment method we use to used comparable information (from the same body parts). This is also the reason why CPS ([Cheng 2011]) provides

the best results. In fact, CPS approach is based on the localization of the body parts, after a heavy learning, and on the extraction and the matching of their descriptors. The fact that our method and [Cheng 2011] one provide the best results demonstrate the importance of precise spacial matching between body part for the re-identification task.

Approach	1st rank (%)	5th rank (%)	nAUC (%)
Our approach	52.94	72.27	93.49
SDALF [Farenzena 2010]	50.62	70.08	93.14
CPS [Cheng 2011]	47.42	73.15	93.52
LCP [Bak 2011]	37.81	54.24	–
Group-Context [Zheng 2009]	23.95	43.26	–
HPE [Bazzani 2010]	21.66	46.76	–

Table 7.2: Detailed results on iLids-119 dataset.

■ **i-LIDS-MA**

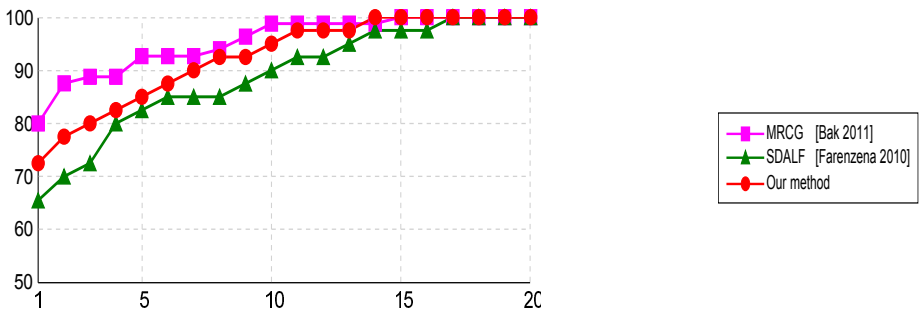


Figure 7.27: CMC curves obtained on iLids-MA dataset.

The results of the evaluation of our approach on the i-LIDS-MA dataset, displayed in figure 7.27 and Table 7.3, shows that our method outperforms the initial one proposed by [Farenzena 2010], but it is still less efficient in comparison with [Bak 2011] approach.

The i-LIDS-MA dataset contains 40 individuals extracted **manually** from two cameras. The individuals are observed from back and back-right sides in both cameras and are well delimited and centred in the images.

The back side acquisition of people in both cameras allows the use of SIFT features and RHSP with covariance descriptors in our signature computation and comparison (their weights are not null), improving the re-identification performances in

comparison with [Farenzena 2010] approach. On the other hand, the textures on people cloth are not sufficient to allow the SIFT and RHSP features to provides the required discriminative power to outperform the "Mean Riemannian Covariance Grid" (MRCG) of [Bak 2011]. The MRCG signature is discriminative in general, and it is especially the case in this dataset, thanks to the good delimitation and alignment of people in the provided images, and thanks to the same side acquisition of all individuals in both cameras.

Note that our approach is significantly faster than [Bak 2011] MRCG which uses 11×11 covariance matrices. In fact, the computation of one person signature using 46 images requires about 6 s with MRCG approach while our method requires about 620 ms for the same 46 images and 138 ms in average using 10 sampled images by counting the computation of SIFT features which are not provided by mono-camera tracking algorithm in this case). This is due to the high computational cost of covariance mean (iterative gradient descent, with all the required eigenvalue decomposition), performed for a large number of grid cells, in comparison with the simple 7×7 covariance matrices computation for RHSP characterization, without any mean covariance computation.

The processing time for color histogram is negligible in comparison with all other feature computation (a simple image browsing with the increase of histogram bins). The Maximally Stable Color Regions (MSCR) computation time for most images is decreased thanks to the apriori information concerning the number of required iterations and the initial clustering parameters, provided by the MCSR computation results on the first images of each visible side class. Finally, if we consider the SIFT feature computation as a part of the visual signature computation (which is a special case here, because it is supposed to be provided by mono-camera tracking algorithm in the whole system use), the SIFT feature detection and selection does not require important processing time since they are performed on the last image of each visible side class, using the grid subdivision as explained in the mono-camera tracking algorithm chapter (see sec. 5.2.2).

Approach	**1st rank (%)**	**5th rank (%)**	**nAUC (%)**
MRCG [Bak 2011]	79.98	92.66	86.01
Our approach	72.50	85.00	83.69
SDALF [Farenzena 2010]	65.50	82.50	80.84

Table 7.3: Detailed results on iLids-MA dataset.

■ **i-LIDS-AA**

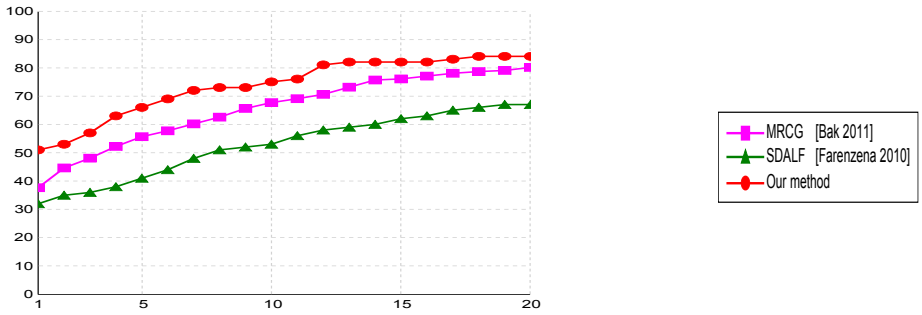


Figure 7.28: CMC curves obtained on iLids-AA datasets.

We have evaluated our approach on the iLids-AA dataset which contains 100 individuals automatically detected and tracked in two cameras. The results are shown in figure 7.28 and Table 7.4. Our approach outperforms both [Farenzena 2010] and [Bak 2011] approaches.

As in iLids-MA dataset, the people images are acquired from the same/close sides (back and back-right sides). This allows our method to use SIFT features and covariance descriptors for RHPS characterization in the visual signature (non null weights), improving the discriminative power of the initial approach of [Farenzena 2010] and outperforming it.

Unlike for iLids-MA dataset, our approach outperforms [Bak 2011] MRCG on this dataset. The main reason is the better management of the badly aligned and delimited images of people by our approach in comparison to [Bak 2011]. Cropped images are noisy and many person images are not centred on the croped images or are missing some parts. The MRCG approach being based on a mean covariance computation using a grid subdivision of images, badly aligned images alter the resulting means. [Bak 2011] proposes a method to deal with this issue by testing grid shift, but the magnitude of the shifting is not sufficient in all situations and increases greatly processing time.

Approach	1st rank (%)	5th rank (%)	nAUC (%)
Our approach	51.00	66.00	94.35
MRCG [Bak 2011]	37.67	55.67	92.87
SDALF [Farenzena 2010]	32.00	41.00	89.29

Table 7.4: Detailed results on iLids-AA dataset.

■ **i-LIDS-AA-RP**

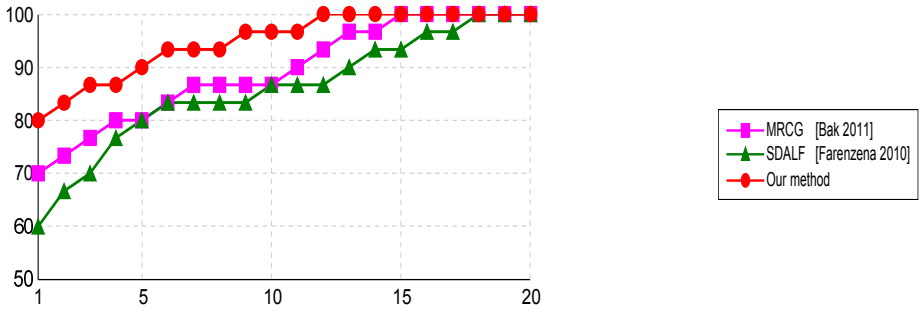


Figure 7.29: CMC curves obtained on iLids-AA-RP datasets.

The results of the evaluation of our approach on the i-LIDS-AA-RP dataset is displayed in figure 7.29 and Table 7.5. We can observe that our approach outperforms both [Farenzena 2010] and [Bak 2011] approaches.

In this dataset, people images do not suffer a lot from bad alignement and delimitation issues like in iLids-AA dataset. This is due to two main reasons: First, we have used our own people detector (chapter 4) and mono-camera tracking (chapter 5) algorithms which outperform the used ones for iLids-AA extraction (based on HOG detector and tracker). Second, we have selected the people following the main criteria that they change their visible side by turning with at least 90° (some people are operating rotations higher than 180°). The chance has made this people fully visible (it was not intentional as long as we were focusing on the visible side changes).

With this dataset, we were expecting [Bak 2011] MRCG to provide the best results, due to the high similarity of this dataset with the iLids-MA one in terms of good people delimitation and alignment in images and the low number of people in the dataset (30 in this dataset and 40 in iLids-MA). The evaluation shows that our approach outperforms [Bak 2011] approach, which has decreased in terms of performances for the first ranks in comparison with iLids-MA even if the number of candidates is lower (the smaller the dataset is, the better the results are expected to be). In fact, [Bak 2011] provides 80% of correct matches at the first rank in iLids-MA while it provides 70% at the first rank for iLids-AA-RP dataset. This shows the dependence of their approach to the visible side of people. The computed mean covariances are altered due to the variation of visible information in the same grid location when a given person turns. Our approach handles better this situation and is less dependent on the orientation of people.

Approach	1st rank (%)	5th rank (%)	nAUC (%)
Our approach	80.00	90.00	80.76
MRCG [Bak 2011]	70.00	80.00	74.96
SDALF [Farenzena 2010]	60.00	80.00	71.36

Table 7.5: Detailed results on iLids-AA-RP dataset.

7.3.3.3 ETHZ Dataset

We have evaluated our approach on the ETHZ dataset too. The results are shown in figure 7.30 and Tables 7.6, 7.7 and 7.8. We can observe that all state of the art approaches perform well on this dataset. The ETHZ dataset is the less challenging one in comparison with the other used datasets, due to the image acquisition system. In fact, all images are acquired using a single camera, avoiding most of the main challenges which can be encountered for real camera network (different color rendering, image resolution, etc.). This, in addition to the low number of people (in comparison with VIPeR dataset for example), explains the good results of our approach and those of the state of the art approaches.

The best results are obtained by CPS ([Cheng 2011]), based on precise body part detection and color histograms, thanks to the good quality of images in this dataset (good resolution and highly rich in terms of color and textures). This image quality allows our approach and LCP one ([Bak 2011]) to have high performances, even if they are slightly worse than CPS approach.

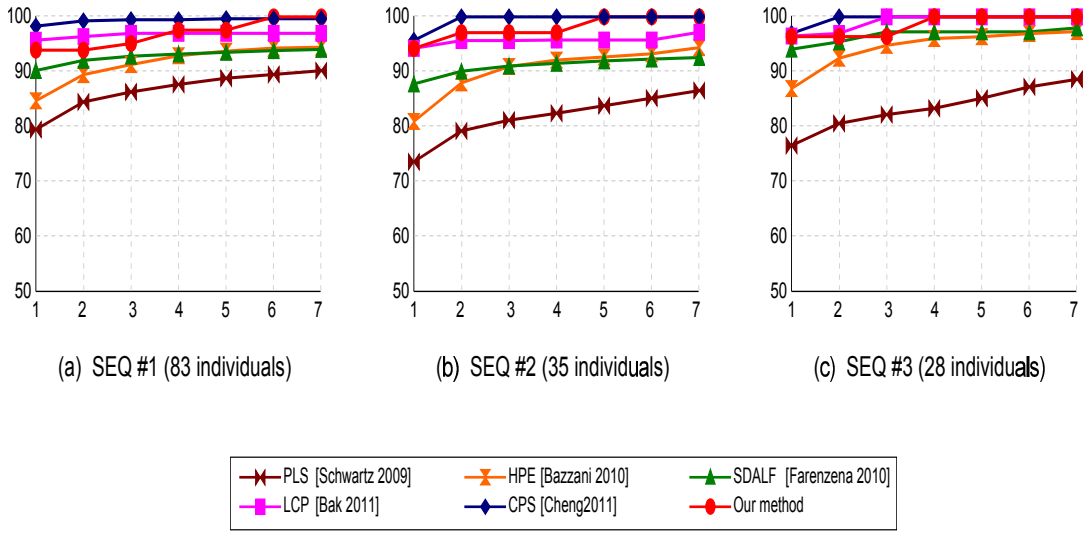


Figure 7.30: CMC curves obtained on ETHZ datasets.

Approach	1st rank (%)	5th rank (%)	nAUC (%)
CPS [Cheng 2011]	98.32	99.66	–
LCP [Bak 2011]	95.78	97.02	–
Our approach	93.98	97.59	99.97
SDALF [Farenzena 2010]	90.29	93.58	98.76
HPE [Bazzani 2010]	84.81	93.80	–
PLS [Schwartz 2009]	79.61	88.90	–

Table 7.6: Detailed results on ETHZ Sequence# 1 (83 individuals).

Approach	1st rank (%)	5th rank (%)	nAUC (%)
CPS [Cheng 2011]	95.74	100	99.93
LCP [Bak 2011]	94.30	95.81	–
Our approach	94.29	100	99.65
SDALF [Farenzena 2010]	87.86	92.03	95.85
HPE [Bazzani 2010]	81.07	92.74	–
PLS [Schwartz 2009]	73.77	83.92	–

Table 7.7: Detailed results on ETHZ Sequence# 2 (35 individuals).

Approach	1st rank (%)	5th rank (%)	nAUC (%)
CPS [Cheng 2011]	97.00	100	99.94
LCP [Bak 2011]	96.53	100	99.82
Our approach	96.43	100	99.67
SDALF [Farenzena 2010]	94.15	97.29	98.96
HPE [Bazzani 2010]	87.02	96.41	–
PLS [Schwartz 2009]	76.71	85.27	–

Table 7.8: Detailed results on ETHZ Sequence# 3 (28 individuals).

7.3.3.4 CAVIAR4REID Dataset

For CAVIAR4REID dataset, we have evaluated two versions of our approach and compared them with state of the art approaches. Initially, this dataset is provided as cropped images of people, without the real world information (people positions in the scene). We have then evaluated a first version of our approach without any real world information, only using appearance based method with the related improvements (geometric body subdivision, SIFT, Covariance descriptors for RHSP characterization and adaptive weighting). For the second version, we have extracted the same set of people using our people detector (chapter 4) and mono-camera tracking (chapter 5) algorithms. We have

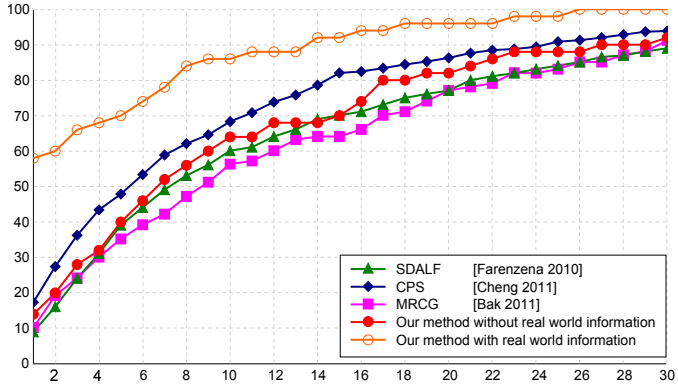


Figure 7.31: CMC curves obtained on CAVIAR4REID dataset.

manually made sure that the same people are available in our extracted dataset and we have selected 10 images per person by sampling all the available images using constant sampling step. This is done to compare with the other approaches. This dataset contains additional real world information for re-identification task and allows us to have global evaluation of the whole system (people detection, mono-camera tracking and people re-identification).

The extremely low resolution of images of the corridor camera makes this dataset very challenging. As expected, our first version (without real world information) outperforms slightly the initial approach of [Farenzena 2010] which outperforms [Bak 2011] MRCG. We suppose that the low improvement of our approach in comparison with [Farenzena 2010] approach is due to the difference between the used weights in our approach and those used in [Farenzena 2010] approach. SIFT features and covariance descriptors for RHSP characterization do not contribute due to the difference of 90° between the two views. Fortunately, the low resolution of the images lowers the entropy of images, and thereby, the importance given to texture, leading to low weights for SIFT and RHSP features. We believe that otherwise, an important weight for SIFT or RHSP features would have decreased the re-identification performances of our approach due to the unavailability of matching between these features from one camera to the other one (perpendicular view axes). Note that this last point is not a real issue in real case application because the SIFT and RHSP weights would be decreased or avoided thanks to the visible side classification stage using real world information. This is demonstrated with the second version of our approach.

Our first version is outperformed by CPS [Cheng 2011] approach, which provides the best results on this dataset (if we considered only pure appearance-base approaches).

In the second version of our approach, we use the available real world information obtained from mono-camera tracking algorithm. As expected, the performances are

significantly better, outperforming CPS [Cheng 2011] approach and providing the best re-identification performances. On the other hand, and surprisingly, the performances are lower than expected. The overlapping field of view between cameras is supposed to help strongly the re-identification process by filtering the incoherent spatio-temporal matches, but it seems that the successive algorithms (people detection and mono-camera tracking), in addition to the equal weighting between real world distance and visual signature matching scores (see eq. 6.17) when many people are too close to each other, have introduced some errors which impact the final re-identification performances.

Approach	**1st rank (%)**	**5th rank (%)**	**nAUC (%)**
Our approach (Whole framework)	58.00	70.00	91.46
CPS [Cheng 2011]	17.36	47.88	82.99
Our approach (initial CAVIAR4REID dataset)	14.00	40.00	79.49
MRCG [Bak 2011]	10.22	35.22	76.20
SDALF [Farenzena 2010]	9.00	39.11	76.24

Table 7.9: Detailed results on CAVIAR4REID dataset.

7.4 Conclusion

In this chapter, we have presented the experimental results of the three proposed algorithms for people detection, mono-camera object tracking and people re-identification in a camera network. For each part, we have presented the usually used metrics and some benchmarking datasets.

7.4.1 People Detector

We have tested various challenging datasets, providing a large amount of complex situations like various scenes, content, and people appearance and poses (INRIA Person dataset), small people images, low resolution (DaimlerChrysler and CAVIAR datasets), low contrast (DaimlerChrysler dataset) and large amount of partial occlusions (Caltech dataset). In addition to the inability to use background subtraction and camera calibration for our people detector in most of these datasets (except for CAVIAR dataset) due to the single images nature or the moving camera acquisition system.

The comparison result shows that the proposed detector have close performances to those of the best detectors of the state of the art. The remaining gap with the best results and with even better results is due to two main issues of our detector: First, the used

covariance descriptors, even if they are very discriminant for regions with small dimensions (from 6×6 in our experiments on INRIA and Caltech dataset), smaller regions in the case of low resolution/noisy images alter the efficiency of this descriptor (observed especially in DaimlerChrysler dataset). The Second issue is related to the full-body type of our detector. Even if the training set of INRIA dataset and in a greater extent the one of Caltech dataset contains people with several levels of partial occlusions, the most miss-detected people in our experiments are due to the partially occluded people, especially for Caltech dataset. Body parts based detectors seems to be more adequate for this kind of situations.

Note that we have shown that people detectors are strongly dependent on the training dataset. In our experiments, we have seen that a trained detector on a given dataset is not applicable on the other datasets with acceptable performances. On one hand, images of the same dataset contain similar characteristics: the same sensor for all acquisitions in a given dataset provides the same resolution and the same noise level, and the similar acquisition conditions for the same dataset provides close people sizes and close point of view acquisition (people pictures acquired by a pedestrian in cities, walking people on sidewalk acquired by embedded camera on a car, etc.). This specializes the trained detector and enables it to integrate these characteristics in the associate thresholds and parameters. On the other hand, these acquisition characteristics differ strongly from a dataset to another one, providing bad performances when detectors are used on different datasets than those used for their training. We suppose that this is a general issue for all people detector approaches since we did not find any evaluation of a unique detector on cross datasets in the literature. For this reason, it is unfortunately not possible to quantify the dependency of the other state of the art detectors to their training datasets.

The dataset dependency of people detectors complicates strongly the wanted genericity for our system (video-surveillance conditions).

7.4.2 Mono-camera Tracking

We have evaluated our mono-camera object tracking algorithm on several sequences from different datasets, in indoor and outdoor situations, with many object occlusion, low object contrast and low object resolution conditions.

The results show that our tracking algorithm provides the best performances on some sequences and close performances to the best trackers on other sequences. The lower performances occur mainly due to crowded scenes. In fact, to track a given object, our tracker requires to learn on-line some parameters on few frames (correct dynamic models, SIFT feature reliability, real world dimension and velocity, dominant colors). To

allow this learning, the object to track should be correctly separated from other objects (no occlusion) for the few first frames in which it appears. This condition is hardly ensured in crowded environments like subway stations.

The state of the art approaches which outperform slightly our tracker use off-line learning to extract the most discriminative features and/or the best parameters for their algorithm. Off-line learning step, even if it improves on-line tracking performances, cannot be envisaged for our tracker due to the deployment requirements of surveillance systems. It may be hard to ask video operators to acquire some sequences and to apply the off-line learning to fine tune the tracking algorithm for each camera, because this operation may be too long for networks with many cameras, but not only. The other main reason is the fact that video operators are generally not experts.

For this reason, our tracker, even if it does not provide the best results on all datasets, represents a good compromise between performances and genericity/autonomy while being deployed in several environments and context.

7.4.3 People re-identification

Finally, we have evaluated our people re-identification algorithm and compared it with state of the art approaches on several datasets. Our method outperforms the initial approach we have taken as basis on all tested datasets, validating our improvements. It also outperforms many other state of the art approaches on several datasets. We have demonstrated the importance of three issues/solutions:

First, we have seen the importance of a correct part-to-part accumulation/comparison on human body. The most effective approach from the state of the art on the tested datasets (PS and CPS [Cheng 2011]) uses an effective body part segmentation. Our image alignment method, even if it is not as precise as a body-part segmentation, seems to be sufficient to improve the re-identification performances in comparison with approaches which do not manage this issue.

Second, we have shown the importance of obtaining the people visible side information. Unfortunately, the used datasets in the state of the art do not highlight enough this issue (even by providing few numbers of people observed from different sides or by selecting people with similar appearance from all sides when the rotation is considered), while this issue is a frequent one in many environments. The dataset we have collected especially for this issue (iLids-AA-RP) shows the effectiveness of our method to manage these cases, thanks to real-world information provided by our mono-camera tracking algorithm.

Finally, we have demonstrated the contribution of the use of context knowledge by using camera calibration allows to improve re-identification performances, by filtering

candidates with incoherent real world positions.

Note that most approaches which provide the best performances (LMNN-R [Dikmen 2011]) on some datasets (VIPeR) use off-line learning step. As it was mentioned for mono-camera tracking and in the objectives of our work, offline learning steps are not considered as possible stages to perform wide scale surveillance, due to the important required time it requires for a large number of cameras and to the inexperience of most video operators with respect to this task.

8

CONCLUSION AND FUTURE WORK

This thesis presents our work to achieve the final aim which is providing a whole framework for people tracking through camera network. In this chapter, we expose the conclusions we have drawn after conducting this work, beginning by highlighting our main contributions (sec. 8.1.1), followed by their limitations (sec. 8.2), and concluding by perspectives (sec. 8.3) which have to be investigated to improve the performances in the corresponding fields.

8.1 Conclusion

In this thesis, we present our methods for people detection, mono-camera object tracking, and people re-identification for video camera networks. The three main constraints which have guided our work, namely high performances, real-time processing, and genericity/easy deployment in industrial context have been generally respected except for the genericity of people detection part. The genericity and easy deployment constraints have led to slightly lower performances in comparison of state of the art approaches in some cases (not always) but the gap between our algorithm performances and the best ones is sufficiently low to be considered as a good compromise between all the constraints.

Wide scale video surveillance constraints and challenges have been addressed and most of them have been well managed.

8.1.1 Contributions

This thesis brings the following three main contributions: an efficient (fast training, real-time/pseudo-real-time detection, high performances) people detector, a robust (high performances), real-time and generic mono-camera object tracking algorithm (turnkey system, without operator configuration or offline learning steps), and a fast (real-time/pseudo-real-time) and generic people re-identification algorithm (turnkey system, without operator configuration or offline learning steps).

The 10 detailed contributions of the presented work in this thesis consists of:

8.1.1.1 An Optimization Method to Improve Cascade of Classifiers for People Detectors

This method has been applied on a state of the art approach, improving its performances significantly while it speeds up both training and detection processing time. It consists in clustering negative data and training each cascade level by the largest remaining cluster. This step has three general and one specific benefit effects: firstly, the use of smaller subsets of negative data for each cascade training speeds up significantly the whole training process. Secondly, the clustering stage provides similar negative contents (with respect to the used information/features), specializing each cascade level to reject this type of content. Thirdly, the use of the largest remaining cluster ensures an optimized cascade level, selecting the most rejecting levels at the first positions, speeding up the detection time. Finally, and this is the most specific benefit, this clustering stage allows to reduce the effect of random selection of candidate weak classifiers when testing all possible candidate weak classifiers is not possible in reasonable time.

8.1.1.2 A New Method for SIFT Feature Detection and Selection for Object Tracking

It consists in a more permissive SIFT feature detection in a first stage, providing a larger number of points on object to track, and in a reliability based selection of a subset of these points on a second stage, according to a grid subdivision of the object, the background subtraction and the SIFT feature robustness. This extension provides a good representation of the whole object to track and allows better tracking especially in partial occlusion cases.

8.1.1.3 An Hybrid Particle Weighting Method for SIFT Feature Particle Filtering

In order to deal with extremal cases (low resolution, noise, small object sizes and bad background subtraction results) which cause bad/lost SIFT feature tracking in some situations, the particles of each SIFT feature are weighted using two different pieces of information: the similarity measure of their descriptors and the "estimated" background/foreground state of their location. To deal with possible errors of background subtraction results, a continuous (not binary) weighting method is proposed by estimating the background subtraction quality and assigning a weight according to its reliability.

8.1.1.4 A Data Association Framework for Object Tracking

This step infers object tracking (localisations and trajectories) from SIFT feature tracking, taking into account the on-line learned reliability of tracked SIFT features and their positions, and managing all possible situations: simple isolated objects, appearing/disappearing objects (leaving the scene or being occluded) and grouping/splitting objects. It creates reliable temporal links between object delimitation on each frame, providing the final object trajectories.

8.1.1.5 A Fast Occlusion Management Method

The mono-camera tracking algorithm being dedicated to static and calibrated cameras as it was mentioned in the hypotheses and constraints (in the introduction chapter), it is able to manage occlusions by taking advantage of useful real world information. We use object real dimensions (width and height) and velocities, after learning their variations during object tracking, to estimate matching scores between occluded objects and a candidate appearing ones. This information is used in addition to the matching score of tracked/detected SIFT features and to dominant color descriptors comparison to get the final matching score and to decide whether a new appearing object corresponds to a previously occluded one.

8.1.1.6 Fast Image Alignments Before Signature Computing for Multiple-shot Case

In order to deal with part-to-part correct matching issue for multiple-shot signature computation, a fast image alignment algorithm is proposed. This algorithm is based on fast search of best matching score, using Lab color distance which is the closest color space to human perception. The use of camera calibration information to reduce searching scales speeds up the image alignment stage. Signatures which are computed after image alignment are more discriminative and more efficient.

8.1.1.7 Use of Texture Information in Addition to Color

The characterisation of Recurrent High-Structured Patches (RHSP) features using covariance descriptors, based on both texture and color information, instead of simple color histograms improves the discriminative power of these features. The use of SIFT features as additional texture information with a negligible additional processing time (due to the fact that they are provided by mono-camera object tracking algorithm) increases this discriminative power and provides better re-identification results when these two features are usable. The local nature of both RHSP and SIFT makes them usable only when the same/nearest sides of people are visible in different cameras.

8.1.1.8 Visible Side Classification for More Reliable Signature Comparison

Appearance of people being different according to the their visible side in many cases, a unique visual signature for each person is not adequate. Assigning a signature to each visible side is a better way to identify a person. A visible side classification method is proposed. This method is fast and does not require complex computations. It is based on people real world trajectories subdivision and clustering, thanks to camera calibration information.

8.1.1.9 Spatio-temporal Coherency Filtering Method

Depending on whether cameras have overlapping fields of view or not, two methods using global camera calibration (in the same coordinate system) information and eventually environment maps if they are available are proposed. In the case of overlapping field of view, people move from a camera to another one by crossing the common field of view. The best opportunity to apply re-identification is then at this moment. Using camera calibration and images to world projections, a surrounding perimeter around the person to re-identify is used to filter all candidates who are too far and to weight visual signature matching scores by a real world distance. In the case of non-overlapping field of view, candidates with incoherent localisation (with respect to their velocity and the elapsed time) are filtered out, reducing the candidate number and thereby, increasing re-identification performances a speed.

8.1.1.10 Adaptive Weights for Signature Components

The proposed visual signature consisting of different features, the importance of each of them depends on the nature of used information and the considered people appearance. Off-line training being proscribed as much as possible, an adaptive method

for feature weighting is proposed. This method assigns the weights per person and not per dataset. It takes in account the visible side of people to decide whether local features (SIFT and RHSP) are used or not, and to assign the several weights according to the richness/poverty of colors and textures.

8.2 Limitations

The proposed approaches still suffer from some limitations. Some of these limitations are specific to our proposed methods while others are more general limitations, impacting all state of the art approaches and constituting open issues. This section presents these limitations and the next section provides some ideas to deal with them, and which will be investigated in future work.

8.2.1 People Detection Limitations

Low resolution, noisy images and small people size

The combination of these three challenges demonstrate the limitation of covariance descriptors. Even if covariance descriptor is a powerful way to encode a large amount of information in a single discriminative and robust structure, it seems that it is not adequate for low resolution images. In addition, too small regions (less than 6×6 pixels) on noisy images are not well characterized.

Processing time

Despite the use of integral images to speed up processing, covariance matrix still have heavy computational cost (due to the several required eigenvalue decompositions). Our people detector, as a stand alone process, can perform in pseudo real time (and in real time for scenes with low complexity), but as a part of a more important framework like ours, containing mono-camera tracking and re-identification tasks which are also time consuming processes, it is not conceivable to use our people detector intensively, on each frame of each sequence, without any region targeting system.

Partial occlusions

The proposed people detector is a full-body based algorithm. It is supposed to be faster than equivalent body-part based detectors (using comparable features) due to the lower testing operations during detection and to the absence of a spatial reasoning step to infer whether detected parts correspond to a human, but present the main disadvantage of being less effective in case of partially occluded people.

Dataset/Acquisition dependency

This is a more general issue for people detection task in the state of the art. Even if the proposed approaches (including ours) are efficient and provide good results when they are tested on data belonging to the same dataset/acquisition conditions than those used for their training, these detectors cannot be used everywhere else, or on a specific dataset if they are trained on a particular other dataset. This issue compromises our search for generic solutions.

8.2.2 Mono-camera Object Tracking Limitations

Dependency on detection results

The proposed tracking algorithm aims at creating temporal links between detected/tracked objects on successive frames. So the tracking algorithm needs to be provided by object localisation and delimitations. This task is performed by a collaboration between a state of the art background subtraction algorithm and our people detector. If some objects are not detected or are badly delimited, it impacts strongly the tracking performances.

Crowd/Occlusion object state at the starting of its tracking

The proposed algorithm being designed for generic and turnkey deployment, off-line learning steps are not considered. The algorithm learns some parameters (SIFT features reliability and real world information) on-line during the tracking. To learn correct values and dynamic models for SIFT features, the proposed algorithm requires for each tracked object to be "easily" identified and tracked for the first frames it appears. This does not necessary mean that the algorithm is only dedicated for sterile zone surveillance, it can track correctly people belonging to a group if their separation has been correctly detected by our people detector and if the group has uniform displacement over few frames (no people crossing during these first frames).

Exclusive use of a unique feature: SIFT

Even if SIFT features are known to be highly discriminative and robust against many condition changes, they still are only texture-based features. We believe that the more the object model is rich in terms in information, the better the tracking is, as long as the information is smartly used and allows real-time processing.

Processing time

The proposed tracking algorithm performs in real time for scenes with maximum number of 7 to 10 tracked objects. The grid subdivision and SIFT features selection methods we have used allow to reduce the impact of object size on processing time since the good

spatial repartition of SIFT features allows to take approximatively constant number of SIFT features even if a person has a large size or not in the image. However, our tracker cannot be used for real time processing in crowded scenes.

8.2.3 People Re-identification Limitations

Low resolution images and small people size

As for people detection, covariance descriptors are not adequate to characterize RHSP with small dimensions (on small people images) especially when images are noisy.

Exclusive use of color for people image alignment

The proposed method to align images of the same person in multi-shot case generally provides good results. However, in some cases, when the contrast between people and the background is low, the computed color distances during the alignment process may lead to an erroneous alignment.

Processing time / sampling method for multiple-shot case

The proposed method performs in real time or pseudo real time depending on the number of used images to compute each person's signature. The more images are considered per person, the better the re-identification performances are, but the slower the processing is. Our method to samples the available images of a person, based only on selecting images after constant trajectory intervals and per visible class side may not be optimal. This sampling method does not take into account any visual information. The sampled images may be too similar and do not represent the possible variations in appearance of the considered person.

Fully autonomous people tracking under camera network

As it was shown in the evaluation chapter, the performances of the re-identification task in the state of the art are not good enough to allow a fully-autonomous people tracking under a camera network. This kind of system requires a high re-identification rate at the first rank to be interesting (as long as the aim is to track reliably and unsupervisedly a person in the camera network). State of the art re-identification approaches are not yet mature enough to address this task for generic and wide scale surveillances systems, with many more people than in the used datasets and with a large number of cameras (which increase the combinatorial complexity and the classical issues). More research work remains to be done to reach this objective. However, many other types of application can take advantage of proposed approaches and their current performances.

We can cite the two main examples which are “Assisted tracking” and “People retrieval a-posteriori”. In fact, assisted tracking is a kind of partial-autonomous system which tries to track people through the camera network, but which can be guided and corrected by the video operator by simple actions such us clicking on the person of interest. For the people a-posteriori retrieval, the contribution of re-identification algorithm may be more important by greatly reducing the number of candidates that human operators or enforcement officiers have to manually check, and this, by proposing only people in the 10% first ranks for example.

8.3 Future Work

8.3.1 Short-term Perspectives

8.3.1.1 People Detection

Another kind or part-based people detector for partial occlusion management

We have highlighted the fact that partial occlusion cases are better managed by body part based detectors. We have avoided this kind of approach due to the higher required processing time and to the additional spatial positioning reasoning. Noting that partial occlusions mainly occur uniformly from one of the four sides of a person (upper side, lower side, right side, left side), and instead of training a large number of body part detectors (for head, shoulders, hands, arms, legs, etc.), another way to explore this challenge may be to train only four detectors on the four global cited sides, with different levels of occlusion (for example, upper occlusion which ends at different levels, starting occlusion from waist to knee for lower body part detector, etc.). This possible solution has two main advantages: the number of parts to consider is lower (only 4) and the spatial positioning reasoning is easier due to the considered parts.

Evaluating genericity of people detectors by merging several datasets

To have more generic people detectors, it may be interesting to train the detectors on enlarged datasets containing many datasets. A preprocessing step to resize all images on the same size is necessary, and the selected size has to produce the lowest amount of information loss and image alteration. This approach may show how a given detector is able to integrate images characteristics (as resolution, noise, people size, acquisition point of view, etc.) as full part features in addition to classical features (color, textures, etc.).

8.3.1.2 Mono-camera Object Tracking

Use of additional and "light" features for tracking initialization

To solve our issue of bad parameter learning in the few first frames when complex situation occurs, a possible solution may be to apply another kind of object tracking, based on simple information like color histograms and pseudo-exhaustive search (low processing time requirement), to provide the first positions of a tracked object and to allow the main tracking algorithm to learn the necessary values.

Integrate color information more efficiently in the tracking process

The proposed approach uses color information only for occlusion management purpose, and in a simple way (dominant color descriptor). It may be more efficient to integrate color information in the core tracking algorithm, by computing Color SIFT features or by tracking color patches using similar particle filter.

8.3.1.3 People Re-identification

Better image sampling for multiple-shot case

Instead of sampling multi-shot images only by selecting images after constant trajectory intervals and per visible class side, which may take too similar images, this sampling should be performed using visual information criteria, by maximizing the variance of the considered features on the images to select. This may lead to use more representative images of possible variations in people appearance for signature computation.

Use of texture as additional information for image alignment

The presented image alignment algorithm provides good results most of the time but fails when the people cloth colors are too close to background ones. Adding texture information may solve this issue in a larger extent.

8.3.2 Long-term Perspectives

8.3.2.1 People Detection

Automatic Context Classification

Depending on the used features for people detection, it may be very useful to identify a way to characterize and classify the deployment environments (or datasets) by grouping them in similar context sets. Once this classification is performed, dedicated people detectors may be trained on each class of context. The final aim is to be able to identify to which class a given new deployment environment (or dataset) belongs, and

to automatically select the corresponding detector which is supposed to provide the best results. This may solve the genericity lack of current people detection systems.

Hybrid and hierarchical use of several features

Even if covariance descriptors have proven their effectiveness in most of cases, their required heavy processing time leads us to consider a smarter way to use them. Some simple cases may be easily managed with faster features like LBP or Haar-Like features. We do not believe that the concatenation of features in the same strong classifiers is the best way to use other types of information. Quite the contrary, we believe that training each cascade level using a unique type of features, by taking care of training the first cascade levels using faster features for simple negatives fast rejection, and use more complex features (like covariance descriptor) at the end of the cascade to reject more complex negatives may be a better way to use multiple features approaches. This may be justified by two reasons: generally, most of negative regions in images are not complex cases (road, walls, sky, cars, etc.). Many fast and simple features may easily be used to reject these simple cases. The second reason is related to our desire to specialize each cascade level to reject a given kind of information.

8.3.2.2 Mono-camera Object Tracking

High level, long term on-line tracking controller

As alternative to off-line learning, which cannot be considered for the addressed requirements, it may be interesting to investigate the on-line tracking controller solutions. In fact, in our tracking algorithm, once a decision is taken by our data association framework concerning the temporal links between objects of the previous frame and those of the current frame, this decision is no longer questioned and the tracking continues even if some links are incorrect. Using long-term control to check the coherency of trajectories provided by our tracker may help to correct bad tracks and to update the state of our tracking for better incoming results.

Multiple feature based tracking, with smart and alternating features use

Generally, the more discriminative a given feature is, the more processing time it requires to be computed. The scene complexity is not constant: sometimes people are isolated while in other times the scene is crowed. It is even possible to have both cases in the same scene at the same time, with crowded zones and relatively empty ones. Starting from these two observations, it may be interesting to find a way to have a multi-feature tracker which switches easily from a feature to another one without tracking interruption and to manage this tracker under a high level controller which may predict

incoming complex situations for a given tracked object, thanks to some predictions on object trajectories in the scene, their separating distances, their velocities, etc. For example, it may be interesting when in the same scene, some objects are fastly tracked using simple color histograms due to the fact that they are isolated while other objects are tracked using more complex and discriminative features like SIFT due to their proximity to each other or because they are partially occluded. This leads to a system in which, at the same moment, in the same scene, some objects are tracked with only their color histograms while other are tracked using SIFT features. This kind of approach allows to save some processing time in simple case tracking. This is a different approach than those [Li 2009, Chau 2011] which use off-line learning step to select the best features which may be used later for all objects.

8.3.2.3 People Re-identification

Automatic Context Classification

As for people detection, an automatic context classification method may help to identify which features are more relevant to reach the best re-identification performances. In our case, the feature weights used for each signature component may be more precisely selected knowing more information concerning the deployment environment (or dataset).

Group-context information as additional descriptor

In many surveillance area (airport, train station, etc.), the possible displacement ways are limited, and people move following very frequent trajectories and thereby, constituting more or less constant groups. It would be interesting to explore group-context additional information, in a more sophisticated way than in [Zheng 2009], by adding real world information combined with all other our proposed improvements.

Use object detectors for usual object removing

Many occlusions are due to usual objects which are taken by people. For example, in airport backpacks, luggages and trolleys are frequently occluding people. When the people are well delimited and the occlusion amount is low with respect to people size (small backpack, carried luggage which occlude a small part of a person, etc.), this occlusion affects the visual signature by integrating wrong information, but the effect may be low. On the other side, if the occlusion amount is high, the computed visual signature is highly altered. To deal with this issue, the use of detectors for usual objects may be an interesting solution. If these objects are well delimited also, their corresponding image regions will not be used in the visual signature computing, and may be replaced by

extrapolated information if their zones are directly on the people (like backpack viewed from behind).

BIBLIOGRAPHY

[Abdel-Hakim 2006] Alaa E. Abdel-Hakim and Aly A. Farag. *CSIFT: A SIFT Descriptor with Color Invariant Characteristics*. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06, pages 1978–1983, Washington, DC, USA, 2006. IEEE Computer Society.

[Ahonen 2004] Timo Ahonen, Abdenour Hadid and Matti Pietikainen. *Face Recognition with Local Binary Patterns*. In Computer Vision - ECCV 2004, volume 3021, chapitre 36, pages 469–481. 2004.

[Ali 2001] A. Ali and J.K. Aggarwal. *Segmentation and recognition of continuous human activity*. In Detection and Recognition of Events in Video, 2001. Proceedings. IEEE Workshop on, pages 28–35, 2001.

[Angelova 2008] D. Angelova and L. Mihaylova. *Extended Object Tracking Using Monte Carlo Methods*. Trans. Sig. Proc., vol. 56, no. 2, pages 825–832, February 2008.

[Bak 2010] Slawomir Bak, Etienne Corvee, Francois Bremond and Monique Thonnat. *Person Re-identification Using Spatial Covariance Regions of Human Body Parts*. In Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, pages 435–440, 2010.

[Bak 2011] Slawomir Bak, Etienne Corvee, Francois Bremond and Monique Thonnat. *Boosted human re-identification using Riemannian manifolds*. Image and Vision Computing, August 2011.

[Barrow 1977] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles and H. C. Wolf. *Parametric correspondence and chamfer matching: two new techniques for image matching*. In Proceedings of the 5th international joint conference on Artificial intelligence - Volume 2, pages 659–663, 1977.

[Bartlett 2002] M. Bartlett, J. Movellan and T. Sejnowski. *Face recognition by independent component analysis*. In IEEE Transactions on Neural Networks, volume 13, pages 1450–1464, 2002.

[Bauml 2010] Martin Bauml, Keni Bernardin, Mika Fischer, Hazim Kemal Ekenel and Rainer Stiefelhagen. *Multi-pose Face Recognition for Person Retrieval in Camera Networks*. 2010.

[Bay 2008] Herbert Bay, Andreas Ess, Tinne Tuytelaars and Luc Van Gool. *Speeded-Up Robust Features (SURF)*. Comput. Vis. Image Underst., vol. 110, no. 3, pages 346–359, 2008.

[Bazzani 2010] Loris Bazzani, Marco Cristani, Alessandro Perina, Michela Farenzena and Vittorio Murino. *Multiple-Shot Person Re-identification by HPE Signature*. In Proceedings of the 2010 20th International Conference on Pattern Recognition, pages 1413–1416, 2010.

[Bazzani 2012] Loris Bazzani. *Beyond Multi-target Tracking*. PhD thesis, Verona, ITALY, 2012.

[Belhumeur 1997] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman. *Eigenfaces vs. Fisherfaces: recognition using class specific linear projection*. vol. 19, no. 7, pages 711–720, 1997.

[Belongie 2002] S. Belongie, J. Malik and J. Puzicha. *Shape Matching and Object Recognition Using Shape Contexts*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 4, pages 509–522, 2002.

[Bertozzi 2007] M. Bertozzi, A. Broggi, M. Del Rose, M. Felisa, A. Rakotomamonjy and F. Suard. *A Pedestrian Detector Using Histograms of Oriented Gradients and a Support Vector Machine Classifier*, 2007.

[Bilinski 2009] Piotr Bilinski, François Bremond and Mohamed-Bécha Kaâniche. *MULTIPLE OBJECT TRACKING WITH OCCLUSIONS USING HOG DESCRIPTORS AND MULTI RESOLUTION IMAGES*. In 3rd International Conference on Imaging for Crime Detection and Prevention, London, Royaume-Uni, December 2009.

[Birchfield 2005] Stanley T. Birchfield and Sriram Rangarajan. *Spatiograms versus Histograms for Region-Based Tracking*. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02, CVPR '05, pages 1158–1163, Washington, DC, USA, 2005. IEEE Computer Society.

[Bird 2005] Nathaniel Bird, Osama Masoud, Nikolaos Pananikolopoulos and Aaron Isaacs. *Detection of Loitering Individuals in Public Transportation Areas*. IEEE Transactions on Intelligent Transportation Systems, vol. 6, no. 2, pages 167–177, 2005.

[Bledsoe 1964] W. W. Bledsoe. *The model method in facial recognition*. In Technical Report PRI 15, Panoramic Research, 1964.

[Bosch 2006] Anna Bosch, Andrew Zisserman and Xavier Muñoz. *Scene classification via pLSA*. In In Proc. ECCV, pages 517–530, 2006.

[Bradski 1998] Gary R. Bradski. *Computer Vision Face Tracking For Use in a Perceptual User Interface*, 1998.

[Cai 2008] Yinghao Cai, Kaiqi Huang and Tieniu Tan. *Human appearance matching across multiple non-overlapping cameras*. In ICPR, pages 1–4, 2008.

[Cai 2010] Yinghao Cai, Valtteri Takala and Matti Pietikainen. *Matching Groups of People by Covariance Descriptor*. In Proceedings of the 2010 20th International Conference on Pattern Recognition, ICPR '10, pages 2744–2747, Washington, DC, USA, 2010. IEEE Computer Society.

[Canny 1986] J Canny. *A Computational Approach to Edge Detection*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 8, no. 6, pages 679–698, 1986.

[Capelli 1999] R. Capelli, A. Lumini, D. Maio and D. Maltoni. *Fingerprint classification by directional image partitioning*. In IEEE Transactions on pattern analysis and machine intelligence, volume 21, 1999.

[Chang 2005] Cheng Chang, Rashid Ansari and Ashfaq Khokhar. *Multiple object tracking with kernel particle filter*. In Proceedings, IEEE Conference on Computer Vision and Pattern Recognition 1 (2005) 566 - 573, pages 566–573, 2005.

[Chau 2011] Duc Phu Chau, François Brémond, Monique Thonnat and Etienne Corvée. *Robust Mobile Object Tracking Based on Multiple Feature Similarity and Trajectory Filtering*. CoRR, vol. abs/1106.2695, 2011.

[Chellappa 2007] Rama Chellappa, Amit K. Roy-Chowdhury and Amit Kale. *Identification using Gait and Face*. 2007.

[Chen 2003] Shu-Ching Chen, Mei-Ling Shyu, S. Peeta and Chengcui Zhang. *Learning-based spatio-temporal vehicle tracking and indexing for transportation multime-*

dia database systems. Intelligent Transportation Systems, IEEE Transactions on, vol. 4, no. 3, pages 154–167, 2003.

[Chen 2007a] Y. Chen, G. Liang, K. K. Lee and Y. Xu. *Abnormal behavior detection by multi-svm-based bayesian network*. pages 298–303, 2007.

[Chen 2007b] Yu-Ting Chen and Chu-Song Chen. *A cascade of feed-forward classifiers for fast pedestrian detection*. In Proceedings of the 8th Asian conference on Computer vision - Volume Part I, ACCV'07, pages 905–914, Berlin, Heidelberg, 2007. Springer-Verlag.

[Chen 2009] Jianjun Chen, Suofei Zhang, Zhenyang Wu and Guocheng An. *Mean shift tracking with Kernel Co-Occurrence Matrices*. In Microelectronics Electronics, 2009. PrimeAsia 2009. Asia Pacific Conference on Postgraduate Research in, pages 253–256, 2009.

[Chen 2011] Zhiwen Chen, Jianzhong Cao, Yao Tang and Linao Tang. *Tracking of moving object based on optical flow detection*. In Computer Science and Network Technology (ICCSNT), 2011 International Conference on, volume 2, pages 1096–1099, 2011.

[Cheng 2011] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani and Vittorio Murino. *Custom Pictorial Structures for Re-identification*. In Proc. BMVC, pages 68.1–68.11, 2011. http://dx.doi.org/10.5244/C.25.68.

[Comaniciu 2000] Dorin Comaniciu, Visvanathan Ramesh and Peter Meer. *Real-Time Tracking of Non-Rigid Objects using Mean Shift*, 2000.

[Comaniciu 2002] Dorin Comaniciu, Peter Meer and Senior Member. *Mean shift: A robust approach toward feature space analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pages 603–619, 2002.

[Comaniciu 2003] Dorin Comaniciu, Visvanathan Ramesh and Peter Meer. *Kernel-Based Object Tracking*, 2003.

[Cootes 2001] Timothy F. Cootes, Gareth J. Edwards and Christopher J. Taylor. *Active Appearance Models*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pages 681–685, 2001.

[Corvee 2009] Etienne Corvee and François Bremond. *Combining face detection and people tracking in video sequences*. In The 3rd International Conference on Imaging for Crime Detection and Prevention - ICDP09, page 1, 2009.

[Corvee 2010] Etienne Corvee and Francois Bremond. *Body Parts Detection for People Tracking Using Trees of Histogram of Oriented Gradient Descriptors*. In Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS '10, pages 469–475, Washington, DC, USA, 2010. IEEE Computer Society.

[Cunado 1997] David Cunado, Mark S. Nixon and John N. Carter. *Using Gait as a Biometric, via Phase-weighted Magnitude Spectra*. In Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, pages 95–102, 1997.

[Dalal 2005] Navneet Dalal and Bill Triggs. *Histograms of Oriented Gradients for Human Detection*. In International Conference on Computer Vision & Pattern Recognition, pages 886–893, 2005.

[Daugman 2002] John Daugman. *How Iris Recognition Works*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, pages 21–30, 2002.

[Dikmen 2011] Mert Dikmen, Emre Akbas, Thomas S. Huang and Narendra Ahuja. *Pedestrian recognition with a learned metric*. In Proceedings of the 10th Asian conference on Computer vision - Volume Part IV, pages 501–512, 2011.

[Dollar 2010] Piotr Dollar, Serge Belongie and Pietro Perona. *The Fastest Pedestrian Detector in the West*. In Proceedings of the British Machine Vision Conference, pages 68.1–68.11. BMVA Press, 2010. doi:10.5244/C.24.68.

[Elgammal 2002] A. Elgammal, R. Duraiswami, D. Harwood and L.S. Davis. *Background and foreground modeling using nonparametric kernel density estimation for visual surveillance*. Proceedings of the IEEE, vol. 90, no. 7, pages 1151–1163, 2002.

[Ess 2007] Andreas Ess, Bastian Leibe and Luc Van Gool. *Depth and Appearance for Mobile Scene Analysis*. Computer Vision, IEEE International Conference on, vol. 0, pages 1–8, 2007.

[Everingham 2009] Mark Everingham, Luc Van Gool, C. K. I. Williams, J. Winn and Andrew Zisserman. *The PASCAL Visual Object Classes (VOC) challenge*, 2009.

[Farenzena 2010] M. Farenzena, L. Bazzani, A. Perina, V. Murino and M. Cristani. *Person re-identification by symmetry-driven accumulation of local features*. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 2360–2367, 2010.

[Fazli 2009] S. Fazli, H.M. Pour and H. Bouzari. *Particle Filter Based Object Tracking with Sift and Color Feature*. In Machine Vision, 2009. ICMV '09. Second International Conference on, pages 89–93, 2009.

[Felzenszwalb 2000] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. *Efficient Matching of Pictorial Structures*. In Proc. IEEE Computer Vision and Pattern Recognition Conf., pages 66–73, 2000.

[Felzenszwalb 2010] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan. *Object Detection with Discriminatively Trained Part-Based Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pages 1627–1645, 2010.

[Figueiredo 2000] Mário A. T. Figueiredo and Anil K. Jain. *Unsupervised Learning of Finite Mixture Models*. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 24, pages 381–396, 2000.

[Finlayson 2005] Graham D. Finlayson, Steven D. Hordley, Gerald Schaefer and Gui Yun Tian. *Illuminant and device invariant colour using histogram equalisation.* Pattern Recognition, vol. 38, no. 2, pages 179–190, 2005.

[Forssén 2007] Per-Erik Forssén. *Maximally Stable Colour Regions for Recognition and Matching*. In IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, June 2007. IEEE Computer Society, IEEE.

[Forsyth 1997] D. A. Forsyth and M. M. Fleck. *Body plans*. In Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), CVPR '97, pages 678–, Washington, DC, USA, 1997. IEEE Computer Society.

[Forsyth 2002] David A. Forsyth and Jean Ponce. Computer vision: A modern approach. Prentice Hall Professional Technical Reference, 2002.

[Freund 1995] Yoav Freund and Robert E. Schapire. *A decision-theoretic generalization of on-line learning and an application to boosting*. In Proceedings of the Second European Conference on Computational Learning Theory, EuroCOLT '95, pages 23–37, 1995.

[Friedman 1998] Jerome Friedman, Trevor Hastie and Robert Tibshirani. *Additive Logistic Regression: a Statistical View of Boosting*. Annals of Statistics, vol. 28, page 2000, 1998.

[Gallagher 2008] Andrew C. Gallagher and Tsuhan Chen. *Clothing cosegmentation for recognizing people*. In In Proc. of Conf. on Computer Vision and Pattern Recognition, 2008.

[Galton 1892] Francis Galton. Fingerprints. 1892.

[Gavrila 2004] D. M. Gavrila, J. Giebel and S. Munder. *Vision-based pedestrian detection: The PROTECTOR system*. In In IEEE Intelligent Vehicles Symposium, pages 13–18, 2004.

[Gerónimo 2006] D. Gerónimo, A.D. Sappa, A.M. Lpez and D. Ponsa. *Pedestrian detection using adaboost learning of features and vehicle pitch estimation*. In In Proceedings of the IASTED Int. Conf. on Visualization, Imaging and Image Processing, pages 400–4005, 2006.

[Gerónimo 2007] David Gerónimo, Antonio López, Daniel Ponsa and Angel D. Sappa. *Haar Wavelets and Edge Orientation Histograms for On—Board Pedestrian Detection*. In Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I, IbPRIA '07, pages 418–425, Berlin, Heidelberg, 2007. Springer-Verlag.

[Gerónimo 2009] David Gerónimo. A global approach to vision-based pedestrian detection for advanced driver assistance systems. Centre de Visió per Computador, 2009.

[Gheissari 2006] Niloofar Gheissari, Thomas B. Sebastian and Richard Hartley. *Person Reidentification Using Spatiotemporal Appearance*. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06, pages 1528–1535, Washington, DC, USA, 2006. IEEE Computer Society.

[Gilbert 2006] Andrew Gilbert and Richard Bowden. *Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity*. In Proceedings of the 9th European conference on Computer Vision - Volume Part II, pages 125–136, 2006.

[Goh 2008] Alvina Goh and René Vidal. *Clustering and dimensionality reduction on Riemannian manifolds*. In CVPR, 2008.

[Gonzalez 2001] Rafael C. Gonzalez and Richard E. Woods. Digital image processing. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd édition, 2001.

[Grauman 2005] Kristen Grauman and Trevor Darrell. *The pyramid match kernel: Discriminative classification with sets of image features*. In In ICCV, pages 1458–1465, 2005.

[Gray 2007] Douglas Gray, S. Brennan and H. Tao. *Evaluating Appearance Models for Recognition, Reacquisition, and Tracking*. In 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), 09/2007 2007.

[Guo 2000] Guodong Guo, Stan Z. Li and Kapluk Chan. *Face Recognition by Support Vector Machines*. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, pages 196–, 2000.

[Gustafsson 2002] Fredrik Gustafsson, Fredrik Gunnarsson, Niclas Bergman, Urban Forssell, Jonas Jansson, Rickard Karlsson and Per-Johan Nordlund. *Particle Filters for Positioning, Navigation and Tracking*, 2002.

[Halici 1996] U. Halici and G. Onguin. *Fingerprint classification through self-organizing feature maps modified to treat uncertainties*. Proceedings of the IEEE, vol. 84, no. 10, 1996.

[Hamdoun 2008] Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu and Bruno Steux. *Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences*. In 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC-08), pages –, Stanford, Palo Alto, États-Unis, 2008.

[He 1990] DC. He and L. Wang. *Texture Unit, Texture Spectrum, And Texture Analysis*. Geoscience and Remote Sensing,IEEE Transactions on, vol. 28, pages 509–512, 1990.

[Heisele 2003] Bernd Heisele, Purdy Ho, Jane Wu and Tomaso Poggio. *Face recognition: component-based versus global approaches*. Comput. Vis. Image Underst., vol. 91, no. 1-2, pages 6–21, 2003.

[Hirzer 2012] Martin Hirzer, Peter M. Roth and Horst Bischof. *Person Re-identification by Efficient Impostor-Based Metric Learning*. In AVSS, pages 203–208. IEEE Computer Society, 2012.

[Hordley 2005] S. D. Hordley, G. D. Finlayson, G. Schaefer and G. Y. Tian. *Illuminant and device invariant colour using histogram equalisation*. Pattern Recognition, vol. 38, page 2005, 2005.

[Horn 1981] Berthold K. P. Horn and Brian G. Schunck. *Determining Optical Flow*, 1981.

[Huang 2008] Chang Huang, Bo Wu and Ramakant Nevatia. *Robust Object Tracking by Hierarchical Association of Detection Responses*. In Proceedings of the 10th European Conference on Computer Vision: Part II, ECCV '08, pages 788–801, Berlin, Heidelberg, 2008. Springer-Verlag.

[Huang 2009] Chung-Hsien Huang, Yi-Ta Wu and Ming-Yu Shih. *Unsupervised Pedestrian Re-identification for Loitering Detection*. In Toshikazu Wada, Fay Huang and Stephen Lin, editeurs, Advances in Image and Video Technology, Third Pacific Rim Symposium, PSIVT 2009, Tokyo, Japan, January 13-16, 2009. Proceedings, volume 5414 of *Lecture Notes in Computer Science*, pages 771–783. Springer, 2009.

[Ijiri 2012] Yoshihisa Ijiri, Shihong Lao, Tony X. Han and Hiroshi Murase. *Human Re-identification through Distance Metric Learning based on Jensen-Shannon Kernel*. In Gabriela Csurka and JosÃ© Braz, editeurs, VISAPP (1), pages 603–612. SciTePress, 2012.

[Iwama 2012] H. Iwama, Y. Makihara and Y. Yagi. *Group Context-aware Person Identification in Video Sequences*. IPSJ Trans. on Computer Vision and Applications, vol. 4, pages 87–99, Jul. 2012.

[Jain 1989] A.K. Jain. Fundamentals of digital image processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[Jain 1997] A.K. Jain, L. Hong, S. Pankanti and R. Bolle. *An identity-authenfication system using fingerprints*. In Proceedings of the IEEE, numéro 9, page 85, 1997.

[Jain 1999] A.K. Jain and S. Pankanti. *FingerCode: a filterbank for fingerprint representati on and matching*. In IEEE CVPR, pages 2187–, 1999.

[Javed 2003] Omar Javed, Zeeshan Rasheed, Khurram Shafique and Mubarak Shah. *Tracking Across Multiple Cameras With Disjoint Views*. In IN THE NINTH IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, pages 952–957, 2003.

[Javed 2005] Omar Javed, Khurram Shafique and Mubarak Shah. *Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras*. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02, CVPR '05, pages 26–33, Washington, DC, USA, 2005. IEEE Computer Society.

[Javed 2008] Omar Javed, Khurram Shafique, Zeeshan Rasheed and Mubarak Shah. *Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views*. Comput. Vis. Image Underst., vol. 109, no. 2, pages 146–162, February 2008.

[Jia 2007] Hui-Xing Jia and Yu-Jin Zhang. *Fast Human Detection by Boosting Histograms of Oriented Gradients*. In Proceedings of the Fourth International Conference on Image and Graphics, ICIG '07, pages 683–688, Washington, DC, USA, 2007. IEEE Computer Society.

[Jojic 2009] Nebojsa Jojic, A. Perina, M. Cristani, V. Murino and B. Frey. *Stel component analysis: Modeling spatial correlations in image class structure*. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 2044–2051, 2009.

[Jones 2003] Michael Jones, Paul Viola, Paul Viola, Michael J. Jones, Daniel Snow and Daniel Snow. *Detecting Pedestrians Using Patterns of Motion and Appearance*. In In ICCV, pages 734–741, 2003.

[Kale 2004] Amit Kale, Aravind Sundaresan, A. N. Rajagopalan, Naresh P. Cuntoor, Amit K. Roy-chowdhury and Volker KrÃ$\frac{1}{4}$ger. *Identification of humans using gait*. IEEE Transactions on Image Processing, vol. 13, pages 1163–1173, 2004.

[Kanade 1973] Takeo Kanade. *Picture Processing System by Computer Complex and Recognition of Human Faces*. In doctoral dissertation, Kyoto University. 1973.

[Kanade 1977] Takeo Kanade. *Computer Recognition of Human Faces*. In Interdisciplinary Sys-tems Research, volume 47, 1977.

[Kang 2004] Jinman Kang, Isaac Cohen and Gerard Medioni. *Object Reacquisition Using Invariant Appearance Model*. In Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4 - Volume 04, pages 759–762, 2004.

[Kanhere 2008] N. K. Kanhere and S. T. Birchfield. *Real-Time Incremental Segmentation and Tracking of Vehicles at Low Camera Angles Using Stable Features*. Trans. Intell. Transport. Sys., vol. 9, no. 1, pages 148–160, March 2008.

[Kelly 1971] Michael David Kelly. *Visual identification of people by computer*. PhD thesis, Stanford, CA, USA, 1971. AAI7112934.

[Kim 2001] Dae Hoon Kim and Jang Soo Ryoo. *Iris identification system and method of identifying a person through iris recognition*, 06 2001.

[Kirby 1990] M. Kirby and L. Sirovich. *Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 1, pages 103–108, January 1990.

[Kragik 2000] D. Kragik and H. I. Christensen. *Tracking Techniques for Visual Servoing Tasks 'Computer Vision and Active Perception Lab*, 2000.

[Kuhn 1955] H. W. Kuhn and Bryn Yaw. *The Hungarian method for the assignment problem*. Naval Res. Logist. Quart, pages 83–97, 1955.

[Kullback 1968] S. Kullback. *Information Theory and Statistics*, 1968.

[Kuo 2010] Cheng-Hao Kuo, Chang Huang and Ram Nevatia. *Multi-target tracking by on-line learned discriminative appearance models.* In CVPR, pages 685–692, 2010.

[Laptev 2006] Ivan Laptev. *Improvements of object detection using boosted histograms*. In In Proc. BMVC, pages 949–958. BMVC, 2006.

[Lee 2003a] Kuang-Chih Lee, J. Ho, Ming-Hsuan Yang and D. Kriegman. *Video-based face recognition using probabilistic appearance manifolds*. vol. 1, pages I–313–I–320, 2003.

[Lee 2003b] L. Lee. *Gait analysis for classification*. 2003.

[Leibe 2005] Bastian Leibe, Edgar Seemann and Bernt Schiele. *Pedestrian Detection in Crowded Scenes*. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, pages 878–885, 2005.

[Levi 2004] Kobi Levi and Yair Weiss. *Learning object detection from a small number of examples: the importance of good features*. In Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition, CVPR'04, pages 53–60, 2004.

[Li 2009] Yuan Li, Chang Huang and R. Nevatia. *Learning to associate: HybridBoosted multi-target tracker for crowded scene*. pages 2953–2960, June 2009.

[Lienhart 2002] Rainer Lienhart and Jochen Maydt. *An Extended Set of Haar-Like Features for Rapid Object Detection*. In IEEE ICIP 2002, pages 900–903, 2002.

[Lin 1997] Shang-Hung Lin, Sun-Yuan Kung and Long-Ji Lin. *Face recognition/detection by probabilistic decision-based neural network*. Trans. Neur. Netw., vol. 8, no. 1, pages 114–132, 1997.

[Lin 2008] Zhe Lin and Larry S. Davis. *Learning Pairwise Dissimilarity Profiles for Appearance Recognition in Visual Surveillance*. In Proceedings of the 4th International Symposium on Advances in Visual Computing, ISVC '08, pages 23–34, Berlin, Heidelberg, 2008. Springer-Verlag.

[Liu 1998] Chengjun Liu and Harry Wechsler. *A Unified Bayesian Framework for Face Recognition*. In ICIP (1), pages 151–155, 1998.

[Liu 2004] Zongyi Liu and Sudeep Sarkar. *Simplest Representation Yet for Gait Recognition: Averaged Silhouette*. In Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04), volume 4, pages 211–214, 2004.

[Lowe 1999] David G. Lowe. *Object Recognition from Local Scale-Invariant Features*. In Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[Lowe 2004] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. Int. J. Comput. Vision, vol. 60, no. 2, pages 91–110, 2004.

[Lucas 1981] Bruce D. Lucas and Takeo Kanade. *An Iterative Image Registration Technique with an Application to Stereo Vision*. pages 674–679, 1981.

[Maini 2009] Raman Maini and Himanshu Aggarwal. *Study and Comparison of Various Image Edge Detection Techniques*. In International Journal of Image Processing (IJIP), volume 3, pages 1–11, 2009.

[Maio 1998] D. Maio and D. Maltoni. *Neural Network Based Minutiae Filtering in Fingerprints*. In Proceedings of the 14th International Conference on Pattern Recognition2, volume 2, pages 1654–, 1998.

[Masek 2003] L. Masek and P. Kovesi. *MATLAB Source Code for a Biometric Identification System based on Iris Patterns*, 2003.

[Matas 2002] J. Matas, O. Chum, M. Urban and T. Pajdla. *Robust Wide Baseline Stereo from Maximally Stable Extremal Regions*. In Proc. BMVC, pages 36.1–36.10, 2002. doi:10.5244/C.16.36.

[McHugh 2009] M. McHugh, J. Konrad, V. Saligrama and P.-M. Jodoin. *Foreground-Adaptive Background Subtraction*. IEEE Signal Processing Letters, vol. 16, no. 1, pages 390–393, May 2009.

[Meden 2013] Boris Meden. *Ré-identification de personnes: Application aux rÃ©seaux de caméras Ã champs disjoints*. PhD thesis, Toulouse, FRANCE, 2013.

[Melo 2006] José Melo, Andrew Naftel, Re Bernardino and José Santos-victor. *Detection and Classification of Highway Lanes Using Vehicle Motion Trajectories*, 2006.

[Miezianko 2008] Roland Miezianko and Dragoljub Pokrajac. *People detection in low resolution infrared videos*. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 0, pages 1–6, 2008.

[Mika 1999] S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K. R. Mullers. *Fisher discriminant analysis with kernels*. Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop, pages 41–48, 1999.

[Mikolajczyk 2004] Krystian Mikolajczyk, Cordelia Schmid and Andrew Zisserman. *Human detection based on a probabilistic assembly of robust part detectors*. In European Conference on Computer Vision, ECCV 2004, May, 2004, volume 3021, pages 69–81, 2004.

[Miyazawa 2005] Kazuyuki Miyazawa and Takafumi Aoki. *A Phase-Based Iris Recognition Algorithm*. In ICB 2006, LNCS 3832, Springer-Verlag, pages 356–365, 2005.

[Monteiro 2007] G. Monteiro, P. Peixoto, and U. Nunes. *Vision-based pedestrian detection using haar-like features*. In Robótica, 2007.

[Mörwald 2009] T. Mörwald, M. Zillich, and M. Vincze. *Edge Tracking of Textured Objects with a Recursive Particle Filter*. In 19th International Conference on Computer Graphics and Vision (Graphicon), 2009.

[Mu 2008] Y. D. Mu, S. C. Yan, Y. Liu, T. Huang and B. F. Zhou. *Discriminative local binary patterns for human detection in personal album*. In Proc. IEEE Computer Vision and Pattern Recognition, pages 1–8, 2008.

[Munder 2006] S. Munder and D. M. Gavrila. *An Experimental Study on Pedestrian Classification*. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 28, no. 11, pages 1863–1868, 2006.

[Murshed 2011] M. Murshed, M.H. Kabir and O. Chae. *Moving object tracking an edge segment based approach*. In International Journal of Innovative Computing, Information and Control, volume 7, pages 3963–3979, 2011.

[Nakajima 2003] Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele and Tomaso Poggio. *Full-body person recognition system*. Pattern Recognition, vol. 36, no. 9, pages 1997–2006, September 2003.

[Nefian 1998] Ara V. Nefian, Monson H. Hayes and III. *Face Detection and Recognition Using Hidden Markov Models*. In in International Conference on Image Processing, pages 141–145, 1998.

[Nghiem 2007] Anh-Tuan Nghiem, François Bremond, Monique Thonnat and Valery Valentin. *ETISEO, performance evaluation for video surveillance systems*. In IEEE International Conference on Advanced Video and Signal based Surveillance, 2007.

[Niyogi 1993] Niyogi and Adelson. *Analyzing and recognizing walking figures in XYT*. Rapport technique 223, 1993.

[Ojala 1996] Timo Ojala, Matti Pietikäinen and David Harwood. *A comparative study of texture measures with classification based on featured distributions*. Pattern Recognition, vol. 29, pages 51–59, 1996.

[Ojala 2002] Timo Ojala, Matti Pietikäinen and Topi Mäenpää. *Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pages 971–987, July 2002.

[Oren 1997] Michael Oren, Constantine Papageorgiou, Pawan Sinha, Edgar Osuna and Tomaso Poggio. *Pedestrian Detection Using Wavelet Templates*. In Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), CVPR '97, pages 193–, 1997.

[Pahlavan 1992] K. Pahlavan and J.O. Eklundh. *A head-eye system-analysis and design*. In CVGIP: Image Understanding, vol. 56, 1992.

[Papageorgiou 1998] Constantine P. Papageorgiou, Michael Oren and Tomaso Poggio. *A General Framework for Object Detection*. In Proceedings of the Sixth International Conference on Computer Vision, 1998.

[Papageorgiou 2000] Constantine Papageorgiou and Tomaso Poggio. *A Trainable System for Object Detection*. Int. J. Comput. Vision, vol. 38, no. 1, pages 15–33, 2000.

[Park 2006] U. Park, A. K. Jain, I. Kitahara, K. Kogure and N. Hagita. *ViSE: Visual Search Engine Using Multiple Networked Cameras*. Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol. 3, pages 1204–1207, 2006.

[Penev 1996] P. Penev and J. Atick. *Local features analysis : A general statistical theory for object representation*. Neural Systems, vol. 7, no. 3, pages 477–500, 1996.

[Pennec 2006] Xavier Pennec, Pierre Fillard and Nicholas Ayache. *A Riemannian Framework for Tensor Computing*. INTERNATIONAL JOURNAL OF COMPUTER VISION, vol. 66, pages 41–66, 2006.

[Pérez 2002] P. Pérez, C. Hue, J. Vermaak and M. Gangnet. *Color-based probabilistic tracking*. In In Proc. ECCV, pages 661–675, 2002.

[Perlibakas 2005] V. Perlibakas. *Face recognition using Principal Component Analysis and Log-Gabor Filters*. In Computer Vision and Pattern Recognition (CVPR), 2005.

[Porikli 2003] Fatih Murat Porikli. *Inter-camera color calibration by correlation model function.* In ICIP (2), pages 133–136, 2003.

[Prosser 2008] B. Prosser, S. Gong and T. Xiang. *Multi-camera Matching using Bi-Directional Cumulative Brightness Transfer Functions*. In Proceedings of the British Machine Vision Conference, pages 64.1–64.10. BMVA Press, 2008. doi:10.5244/C.22.64.

[Qian 2007] Huimin Qian, Yaobin Mao, Jason Geng and Zhiquan Wang. *Object tracking with self-updating tracking window*. In Proceedings of the 2007 Pacific Asia conference on Intelligence and security informatics, PAISI'07, pages 82–93, Berlin, Heidelberg, 2007. Springer-Verlag.

[Ramanan 2003] Deva Ramanan and D. A. Forsyth. *Finding and Tracking People from the Bottom Up*, 2003.

[Ronfard 2002] Remi Ronfard, Cordelia Schmid and Bill Triggs. *Learning to parse pictures of people*. In In European Conference on Computer Vision, pages 700–714, 2002.

[Rosten 2006] Edward Rosten and Tom Drummond. *Machine learning for high-speed corner detection*. In Proceedings of the 9th European conference on Computer Vision - Volume Part I, ECCV'06, pages 430–443, Berlin, Heidelberg, 2006. Springer-Verlag.

[Schölkopf 1998] Bernhard Schölkopf, Alexander Smola and Klaus-Robert Müller. *Nonlinear component analysis as a kernel eigenvalue problem*. Neural Comput., vol. 10, no. 5, pages 1299–1319, 1998.

[Schwartz 2009] William Robson Schwartz and Larry S. Davis. *Learning Discriminative Appearance-Based Models Using Partial Least Squares*. In Proceedings of the 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing, pages 322–329, 2009.

[Serby 2004] David Serby and Luc Van Gool. *Probabilistic object tracking using multiple features*. In In IEEE International Conference of Pattern Recognition (ICPR, pages 184–187, 2004.

[Shin 2005] Jeongho Shin, Sangjin Kim, Sangkyu Kang, Seong-Won Lee, Joonki Paik, Besma Abidi and Mongi Abidi. *Optical flow-based real-time object tracking using non-prior training active feature model*. Real-Time Imaging, vol. 11, no. 3, pages 204–218, June 2005.

[Simard 1999] Patrice Y. Simard, Léon Bottou, Patrick Haffner and Yann LeCun. *Boxlets: a fast convolution algorithm for signal processing and neural networks*. In Proceedings of the 1998 conference on Advances in neural information processing systems II, pages 571–577, 1999.

[Sirovich 1987] L. Sirovich and M. Kirby. *Low-dimensional procedure for the characterization of human faces*. J. Opt. Soc. Am. A, vol. 4, pages 519–524, 1987.

[Smith 1992] A. F. M. Smith and A. E. Gelfand. *Bayesian Statistics without Tears: A Sampling-Resampling Perspective*. The American Statistician, vol. 46, no. 2, pages 84–88, 1992.

[Søndrål 2005] T. Søndrål. Using the human gait for authentication. Master's thesis, Gjøvik University College, 2005.

[Sonka 2007] Milan Sonka, Vaclav Hlavac and Roger Boyle. Image processing, analysis, and machine vision. Thomson-Engineering, 2007.

[Souded 2011] Malik Souded, Laurent Giulieri and Francois Bremond. *An Object Tracking in Particle Filtering and Data Association Framework, Using SIFT Features*. In International Conference on Imaging for Crime Detection and Prevention (ICDP), London, UK, 2011.

[Souded 2013] Malik Souded and François Bremond. *Optimized Cascade of Classifiers for People Detection Using Covariance Features*. In International Conference on Computer Vision Theory and Applications (VISAPP), Barcelona, Spain, 2013.

[Stauffer 1999] Chris Stauffer and W. E L Grimson. *Adaptive background mixture models for real-time tracking*. In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., volume 2, pages –252 Vol. 2, 1999.

[Tanner 1987] M Tanner and W Wong. *The calculation of posterior distributions by Data Augmentation (with discussion)*. Journal of the American Statistical Association, vol. 82, pages 528–550, 1987.

[Thayananthan 2003] A. Thayananthan, B. Stenger, P. H. S. Torr and R. Cipolla. *Shape context and chamfer matching in cluttered scenes*. In Proceedings of the 2003 IEEE computer society conference on Computer vision and pattern recognition, CVPR'03, pages 127–133, Washington, DC, USA, 2003. IEEE Computer Society.

[Tipping 2000] Michael E. Tipping. *The Relevance Vector Machine*, 2000.

[Tipping 2001] Michael E. Tipping. *Sparse bayesian learning and the relevance vector machine*. J. Mach. Learn. Res., vol. 1, pages 211–244, September 2001.

[Tissainayagam 2003] Tissainayagam, Suter, P. Tissainayagam and D. Suter. *Object Tracking in Image Sequences using Point Features*. In APRS Workshop on Digital Image Computing Online Proceedings http://www.aprs.org.au/wdic2003/CDROM/69.pdf D. Comaniciu and, pages 1197–1203, 2003.

[Torresani 2006] Lorenzo Torresani and Kuang-Chih Lee. *Large Margin Component Analysis*. In Bernhard SchÃ¶lkopf, John C. Platt and Thomas Hoffman, editeurs, Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006, pages 1385–1392. MIT Press, 2006.

[TranSafety 1997] Inc TranSafety. *Study Compares Older and Younger Pedestrian Walking Speeds*, October 1997.

[Truong Cong 2009] Dung Nghi Truong Cong, Catherine Achard, Louahdi Khoudour and Lounis Douadi. *Video Sequences Association for People Re-identification across Multiple Non-overlapping Cameras*. In Proceedings of the 15th International Conference on Image Analysis and Processing, ICIAP '09, pages 179–189, Berlin, Heidelberg, 2009. Springer-Verlag.

[Truong Cong 2010] D. N. Truong Cong, L. Khoudour, C. Achard, C. Meurie and O. Lezoray. *People re-identification by spectral classification of silhouettes*. Signal Process., vol. 90, no. 8, pages 2362–2374, 2010.

[Turk 1991] Matthew Turk and Alex Pentland. *Eigenfaces for recognition*. J. Cognitive Neuroscience, vol. 3, no. 1, pages 71–86, 1991.

[Tuzel 2006] Oncel Tuzel, Fatih Porikli and Peter Meer. *Region Covariance: A Fast Descriptor for Detection And Classification*. In In Proc. 9th European Conf. on Computer Vision, pages 589–600, 2006.

[Tuzel 2007] Oncel Tuzel, Fatih Porikli and Peter Meer. *Human detection via classification on riemannian manifolds*. In IN PROC. OF THE IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, pages 1–8, 2007.

[Vapnik 1995] Vladimir N. Vapnik. The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[Verma 2011] Abhishek Verma, Chengjun Liu and Jiancheng Jia. *New colour SIFT descriptors for image classification with applications to biometrics*. Int. J. Biometrics, vol. 3, no. 1, pages 56–75, December 2011.

[Viola 2001] Paul Viola and Michael Jones. *Rapid object detection using a boosted cascade of simple features*. pages 511–518, 2001.

[Viola 2003] M. Viola, Michael J. Jones and Paul Viola. *Fast Multi-view Face Detection*. In Proc. of Computer Vision and Pattern Recognition, 2003.

[Viola 2004] Paul Viola and Michael J. Jones. *Robust Real-Time Face Detection*. Int. J. Comput. Vision, vol. 57, no. 2, pages 137–154, May 2004.

[Walk 2010] Stefan Walk, Nikodem Majer, Konrad Schindler and Bernt Schiele. *New Features and Insights for Pedestrian Detection*. In Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[Wang 1990] Li Wang and Dong-Chen He. *Texture classification using texture spectrum*. Pattern Recogn., vol. 23, no. 8, pages 905–910, August 1990.

[Wang 2003a] Liang Wang, Tieniu Tan, Weiming Hu and Huazhong Ning. *Automatic gait recognition based on statistical shape analysis.* vol. 12, no. 9, pages 1120–1131, 2003.

[Wang 2003b] Liang Wang, Tieniu Tan, Huazhong Ning and Weiming Hu. *Silhouette Analysis-Based Gait Recognition for Human Identification*. In IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 25, pages 1505–1518, 2003.

[Wang 2007] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher and Peter Tu. *Shape and Appearance Context Modeling*. Computer Vision, IEEE International Conference on, vol. 0, pages 1–8, 2007.

[Wildes 1997] RICHARD P. Wildes. *Iris recognition: an emerging biometric technology*. Proceedings of The IEEE, vol. 85, pages 1348–1363, 1997.

[Wiskott 1997] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger and Christopher von der Malsburg. *Face Recognition by Elastic Bunch Graph Matching*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pages 775–779, 1997.

[Wren 1997] Christopher Wren, Ali Azarbayejani, Trevor Darrell and Alex Pentland. *Pfinder: Real-Time Tracking of the Human Body*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pages 780–785, 1997.

[Wu 2005] Bo Wu and Ram Nevatia. *Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors*. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01, ICCV '05, pages 90–97, 2005.

[Wu 2007] Bo Wu and Ram Nevatia. *Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors*. Int. J. Comput. Vision, vol. 75, no. 2, pages 247–266, 2007.

[Xing 2009] Junliang Xing, Haizhou Ai and Shihong Lao. *Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses*. 2012 IEEE Conference on Computer Vision and Pattern Recognition, vol. 0, pages 1200–1207, 2009.

[Yang 2005] Changjiang Yang, Ramani Duraiswami and Larry Davis. *Fast multiple object tracking via a hierarchical particle filter*. In In: International Conference on Computer Vision, pages 212–219, 2005.

[Yang 2012] Yi Yang and Deva Ramanan. *Articulated Human Detection with Flexible Mixtures-of-Parts*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, no. PrePrints, page 1, 2012.

[Yao 1995] Yi-Sheng Yao and R. Chellappa. *Tracking a dynamic set of feature points*. Trans. Img. Proc., vol. 4, no. 10, pages 1382–1395, October 1995.

[Yao 2008] Jian Yao and Jean-Marc Odobez. *Fast Human Detection from Videos Using Covariance Features*. In The Eighth International Workshop on Visual Surveillance - VS2008, 2008.

[Yilmaz 2004] A. Yilmaz, Xin Li and M. Shah. *Contour-based object tracking with occlusion handling in video acquired using mobile cameras*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 26, no. 11, pages 1531–1536, 2004.

[Yilmaz 2006] Alper Yilmaz, Omar Javed and Mubarak Shah. *Object tracking: A survey*. ACM Comput. Surv., vol. 38, no. 4, 2006.

[Yoo 2002] J. H. Yoo, M. S. Nixon and C. J. Harris. *Extracting Gait Signatures based on Anatomical Knowledge*. In Proceedings of BMVA Symposium on Advancing Biometric Technologies, 2002.

[Yu 2007] Yang Yu, David Harwood, Kyongil Yoon and Larry S. Davis. *Human appearance modeling for matching across video sequences*. Mach. Vision Appl., vol. 18, no. 3, pages 139–149, 2007.

[Zheng 2009] Wei-Shi Zheng, Shaogang Gong and Tao Xiang. *Associating Groups of People*. In BMVC'09, 2009.

[Zheng 2011] Wei-Shi Zheng, Shaogang Gong and Tao Xiang. *Person re-identification by probabilistic relative distance comparison*. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11, pages 649–656, Washington, DC, USA, 2011. IEEE Computer Society.

[Zhou 009] Huiyu Zhou, Yuan Yuan and Chunmei Shi. *Object tracking using {SIFT} features and mean shift*. Computer Vision and Image Understanding, vol. 113, no. 3, pages 345–352, 2009".

[Zhou 2006] Quming Zhou and J. K. Aggarwal. *Object tracking in an outdoor environment using fusion of features and cameras*. Image Vision Comput., vol. 24, no. 11, pages 1244–1255, 2006.

[Zhou 2012] Shenghui Zhou, Qing Liu, Jianming Guo and Yuanyuan Jiang. *ROI-HOG and LBP based human detection via shape part-templates matching*. In Proceedings of the 19th international conference on Neural Information Processing - Volume Part V, ICONIP'12, pages 109–115, Berlin, Heidelberg, 2012. Springer-Verlag.

[Zhu 2006a] Guopu Zhu, Qingshuang Zeng and Changhong Wang. *Efficient edge-based object tracking*. Pattern Recognition, vol. 39, pages 2223–2226, 2006.

[Zhu 2006b] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng and Shai Avidan. *Fast Human Detection Using a Cascade of Histograms of Oriented Gradients*. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06, pages 1491–1498, Washington, DC, USA, 2006. IEEE Computer Society.