Distributed and higher-order graphical models : towards segmentation, tracking, matching and 3D model inference

Chaohui Wang

To cite this version:

Chaohui Wang. Distributed and higher-order graphical models : towards segmentation, tracking, matching and 3D model inference. Other. Ecole Centrale Paris, 2011. English. NNT : 2011ECAP0037 . tel-00658765

HAL Id: tel-00658765

https://theses.hal.science/tel-00658765

Submitted on 4 Jul 2014

ECOLE CENTRALE PARIS

PHD THESIS

to obtain the title of

Doctor of Ecole Centrale Paris
Specialty : Applied Mathematics

Distributed and Higher-Order Graphical Models:

towards Segmentation, Tracking, Matching and 3D Model Inference

Defended by

Chaohui Wang

prepared at Ecole Centrale Paris, MAS laboratory

defended on September 29, 2011

JURY

Chairman :	Prof. Henri Maitre	-	Télécom ParisTech
Reviewers :	Prof. Michael J. Black	-	Max Planck Institute for Intelligent Systems
	Prof. Philip H. S. Torr	-	Oxford Brookes University
Advisor :	Prof. Nikos Paragios	-	Ecole Centrale Paris
Examiners :	Prof. Patrick Bouthemy	-	INRIA - Rennes
	Prof. Vladimir Kolmogorov	-	Institute of Science and Technology Austria
	Prof. Dimitris Samaras	-	Stony Brook University

Abstract

This thesis is devoted to the development of graph-based methods that address several of the most fundamental computer vision problems, such as segmentation, tracking, shape matching and 3D model inference.

The first contribution of this thesis is a unified, single-shot optimization framework for simultaneous segmentation, depth ordering and multi-object tracking from monocular video sequences using a pairwise Markov Random Field (MRF). This is achieved through a novel 2.5D layered model where object-level and pixel-level representations are seamlessly combined through local constraints. Towards introducing high-level knowledge, such as shape priors, we then studied the problem of non-rigid 3D surface matching. The second contribution of this thesis consists of a higher-order graph matching formulation that encodes various measurements of geometric/appearance similarities and intrinsic deformation errors. As the third contribution of this thesis, higher-order interactions were further considered to build pose-invariant statistical shape priors and were exploited for the development of a novel approach for knowledge-based 3D segmentation in medical imaging which is invariant to the global pose and the initialization of the shape model. The last contribution of this thesis aimed to partially address the influence of camera pose in visual perception. To this end, we introduced a unified paradigm for 3D landmark model inference from monocular 2D images to simultaneously determine both the optimal 3D model and the corresponding 2D projections without explicit estimation of the camera viewpoint, which is also able to deal with misdetections/occlusions.

Keywords: Markov Random Fields, Higher-order MRFs, Segmentation, Tracking, Depth Ordering, Shape Matching, Shape Prior, 3D Model Inference

Résumé

Cette thèse est dédiée au développement de méthodes à base de graphes, permettant de traiter les problèmes fondamentaux de la vision par ordinateur tels que la segmentation, le suivi d'objets, l'appariement de formes et l'inférence de modèles 3D.

La première contribution de cette thèse est une méthode unifiée reposant sur un champ de Markov aléatoire (MRF) d'ordre deux permettant de réaliser en une seule étape la segmentation et le suivi de plusieurs objets observés par une caméra unique, tout en les ordonnançant en fonction de leur distance à la caméra. Nous y parvenons au moyen d'un nouveau modèle stratifié (2.5D) dans lequel une représentation bas-niveau et une représentation haut-niveau sont combinées par le biais de contraintes locales. Afin d'introduire des connaissances de haut niveau a priori, telles que des a priori sur la forme des objets, nous étudions l'appariement non-rigide de surfaces 3D. La seconde contribution de cette thèse consiste en une formulation générique d'appariement de graphes qui met en jeu des potentiels d'ordre supérieur et qui est capable d'intégrer différentes mesures de similarités d'apparence, de similarités géométriques et des pénalisations sur les déformations des formes. En tant que la troisième contribution de cette thèse, nous considérons également des interactions d'ordre supérieur pour proposer un a priori de forme invariant par rapport à la pose des objets, et l'exploitons dans le cadre d'une nouvelle approche de segmentation d'images médicales 3D afin d'obtenir une méthode indépendante de la pose de l'objet d'intérêt et de l'initialisation du modèle de forme. La dernière contribution de cette thèse vise à surmonter l'influence de la pose de la caméra dans les problèmes de vision. Nous introduisons un paradigme unifié permettant d'inférer des modèles 3D à partir d'images 2D monoculaires. Ce paradigme détermine simultanément le modèle 3D optimal et les projections 2D correspondantes sans estimer explicitement le point de vue de la caméra, tout en gérant les mauvaises détections et les occlusions.

Mots-clés : Champs de Markov Aléatoires, Champs de Markov Aléatoires d'ordre supérieur, Segmentation, Suivi, Ordonnancement par Profondeur, Appariement de Formes, A priori de Forme, Inférence de Modèles 3D

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my thesis advisor, Prof. Nikos Paragios, for his insight, guidance and continuous encouragement throughout my PhD studies. He is not only an advisor, but also a role model and a friend to me. After these years of working with him, I have learnt from him to be a researcher, such as how to find and approach interesting computer vision problems, how to write scientific articles and present works orally, how to organize time when dealing with multiple tasks, how to balance work and life, and many others. All these will be beneficial for my future career and research journey.

I am greatly indebted to all my thesis committee members: the reviewers Prof. Michael J. Black and Prof. Philip H. S. Torr, the chairman Prof. Henri Maître, and the examiners Prof. Patrick Bouthemy, Prof. Vladimir Kolmogorov and Prof. Dimitris Samaras, for having taken the time to evaluate my thesis. Their works led me to diverse vision problems and was a considerable source of inspiration for my works. It is my honor and privilege to have their constructive and fruitful comments, which are very valuable not only for this thesis but also for my future research.

I feel very fortunate to have worked with a number of colleagues during my PhD studies. Martin de La Gorce, who was a senior PhD student when I first joined the group of Prof. Paragios, provided enormous help during my early research. He has always been very kind and patient in our numerous discussions, many of which are still vivid in my memory. I also feel very lucky to have known Yun Zeng (*a.k.a.*, Xiang Zeng) during my PhD years. Thanks to many common interests between us, especially those on various computer vision problems, we have had a lot of discussions, which considerably advanced our research. I would also like to deeply acknowledge him for having strengthened my passion for research and also for having encouraged me during tough times. I also have enjoyed collaborating with other colleagues. To name a few, they are Haithem Boussaid, Loïc Simon, Olivier Teboul, Fabrice Michel, Bo Xiang and Alexandros Panagopoulos, *etc*. I would like to thank them for fruitful discussions on various interesting projects.

Many thanks go to Prof. Dimitris Samaras and Prof. Ioannis Kakadiaris for having invited me to visit their labs in 2009 and in 2010, respectively, where I had a great time discussing on vision problems and experiencing American lifestyles. I would like to acknowledge Prof. David Xianfeng Gu and Prof. Michael M. Bronstein for their collaborations and insightful discussions on shape analysis problems. It was a great pleasure working with them. Their deep understanding of geometry gave me much insight and inspiration for my research on related topics. I am also grateful to Prof. Nikos Komodakis and Prof. Iasonas Kokkinos who were always available to answer my questions with their expertise.

I would like to express my sincere gratitude to all my colleagues and friends in the lab for their friendship and company, for providing a great work atmosphere and for helping me in various ways. I have enjoyed and got used to studying and laughing with Loïc Simon, Aristeidis Sotiras and Olivier Teboul, together with whom I have spent five years (starting from our master program) full of interesting memories. I had chance to have worked in the same office with Stavros Alchatzidis, Pierre-Yves Baudin, Haithem Boussaid, Martin de la Gorce, Katerina Gkirtzou, Nicolas Honnorat. We have had a lot of discussions and fun during the past few years. My pleasurable stay at ECP also owes to Regis Behmo, Ahmed Besbes, Salma Essafi, Charlotte Ghys, Radhouene Neji, Daniel Pescia, Mickaël Savinaud, Panagiotis Koutsourakis, Fabrice Michel, Sarah Parisot, Stavros Tsogkas and Bo Xiang. I do not forget Noura Azzabou, Kostas Karantzalos, Olivier Juan, Georg Langs and Maxime Taron, who gave me much useful suggestion to start the PhD studies, though we did not spent much time together. I would also like to thank Olfa Belguith, Pascale Beliveau, Sylvain Berhnardt, Rola Harmouche, Jean Honorio, Wanping Lu, Chris McIntosh, Krishna Nanda, Yangming Ou, Thanos Papadopoulos and Jose Carlos Rubio for having brought a lot of joy to the lab during their short visiting time.

I would like to acknowledge Sylvie Dervin and Annie Glomeron from MAS Laboratory, Lydie Adouvi, Géraldine Carbonel, Catherine Lhopital and Caroline Sbaizero from Office of Research and Graduate Division, and Christine Biard from INRIA Saclay - Ile-de-France for their kind and patient answers to my concerns regarding different administrative processes and for their strong enthusiasm for discussing on interesting subjects. I am also grateful to Prof. Frédéric Abergel, Prof. Paul-Henry Cournède, Prof. Florian De Vuyst, Prof. Gilles Faÿ, Prof. Pascal Laurent and Prof. Jean-Hubert Schmitt, with whom it is always a great pleasure to have discussions.

I would also like to thank numerous friends for their company and support in different ways. Their friendship is a precious lifetime fortune. Particular thanks go to Aline Barazer and Loïc Simon, Mihai si Consuela Elena Jitia Sardarescu, Jing Shen and Haixiang Zhao, who have kindly *offered fuel in snowy weather*.

Finally, I dedicate my thesis to my parents, who continuously encourage me to pursue my interests since my childhood. I'm cordially thankful for their unconditional support throughout my educational life, in particular the past seven years of studies abroad.

Contents

List of Figures

Chapter 1

Introduction

The goal of computer vision is to enable the machine to understand the world - often called *visual perception* - through processing of digital signals. Such an understanding for the machine is done by extracting useful information from the signals and performing complex reasoning. To this end, perception is often associated with the estimation of a set of parameters about the underlying scene, and the inference of these parameters corresponds to the solution of a specific vision problem. Mathematically, let $\mathbf{I}$ denote the observed data (*e.g.*, digital images, surface meshes, *etc*.) and $\mathbf{x}$ denote a latent parameter vector of interest that corresponds to a mathematical answer to the visual perception problem. Perception can be formulated mathematically as finding a mapping from $\mathbf{I}$ to $\mathbf{x}$, which is essentially an *inverse problem* [Szeliski 2010].

Mathematical methods such as variational techniques and statistical methods usually model such a mapping through an optimization problem as follows:

$$\mathbf{x}^{\text{opt}} = \arg\min_{\mathbf{x}} E(\mathbf{x};\mathbf{I}) \tag{1.1}$$

where the energy (or cost, objective) function $E(\mathbf{x};\mathbf{I})$ can be regarded as a quality measure of a parameter configuration $\mathbf{x}$ in the solution space, given the observed images $\mathbf{I}$. Hence, visual perception involves two main tasks: *modeling* and *optimization*. The modeling of a vision problem has to accomplish: (i) the choice of an appropriate representation[1] of the solution using a tuple of variables $\mathbf{x}$; and (ii) the design of the energy function $E(\mathbf{x};\mathbf{I})$ which can correctly measure the adequacy between $\mathbf{x}$ and $\mathbf{I}$. The optimization has to

[1] For example, image segmentation problems can be formulated either as a pixel labeling problem where each variable in $\mathbf{x}$ represents the index of segment for the corresponding pixel or a boundary labeling problem where each binary variable in $\mathbf{x}$ indicates if the boundaries of the segmentation are present at the corresponding edge between a pair of neighbor pixels, *etc*.

search for the set of parameters producing the optimum of the energy function where the solution of the original problem lies.

The main difficulties in the modeling are due to the fact that most of the vision problems are inverse and ill-posed and require a large number of latent and/or observed variables to express the expected variations of the perception answer. Furthermore, the observed signals are usually noisy, incomplete and often only provide a partial view of the desired space. Physics-based, probabilistic and statistical models are often considered to recover latent variables of interest from insufficient observed information. Hence, a successful model usually requires a reasonable *regularization*, a robust *data measure*, and a compact *structure* between the variables of interest to well characterize their relationship (which is usually unknown). In the Bayesian paradigm, the *model prior*, the *data likelihood* and the *dependence properties* correspond respectively to these terms, and the maximization of the posterior probability of the latent variables corresponds to the minimization of the energy function in Eq. 1.1. In addition to these, another issue that should be taken into account during the modeling is the tractability of the optimization task. Such a viewpoint impacts the quality of the obtained optima and introduce additional constraints on the modeling step.

During the past decades, computer vision has made substantial progress thanks to the advance in related fields such as mathematics, statistics, optimization, machine learning, and also to the continuous increase - at a moderate cost - of available computational resources. Numerous mathematical models have been proposed to deal with different vision problems such as image segmentation, tracking and motion analysis, image reconstruction, 3D reconstruction from 2D images and medical image analysis [Paragios *et al.* 2005]. Due to the complexity intrinsically involved in the visual world, more and more researchers have been resorting to a rigorous modeling of physical phenomena, integration of various useful cues/information within a single formulation (*e.g.*, the principled fusion of prior knowledge about objects and data evidence) and/or a joint modeling for complementary tasks (*e.g.*, joint segmentation and tracking), in order to develop more robust algorithms. Compared to early methods such as knowledge-free image segmentation approaches and tackling segmentation and tracking sequentially, such strategies have shown to lead to a better performance and robustness. However, these benefits do not come for free, resulting in a drastic increase in the number of variables (or degrees of freedom) in order to properly treat various tasks in a single formulation. Many existing methods that belong to the scope of variational techniques or statistical methods are based on "*global models*". In such a context, all the variables are coupled such that the objective function cannot be decomposed/factorized. Despite their mathematical soundness, such methods pose challenging issues to the optimization process, since the objective function is in general non-linear, high-dimensional and non-convex with numerous local minima. An even more challenging case is the one where both continuous and discrete variables are present in the objective function. Due to all these facts, many methods resort to coordinate-descent or Expectation-Maximization (EM) optimization approaches to search for the optimal solution. However, it is generally admitted that such optimization schemes are prone to be trapped in local minima and provide no guarantee on the optimality of the solution, which often prevents us from exploring the full expressiveness of the model.

Graph-based approaches - such as Markov Random Fields (MRFs) - that have benefited from recent development in discrete optimization, refer to a promising methodology for solving various vision problems. Such methods provide an excellent compromise between the expressive power of the modeling process and the optimality properties of the corresponding inference algorithms. First, graphical models refer to a modular, flexible and principled way to combine regularization (or prior), data likelihood terms and other useful cues within a single graph-formulation, where continuous and discrete variables can be simultaneously considered. The use of graph provides a simple way to visualize the structure of a model and facilitates the choice and design of the model. Furthermore, the use of discrete optimization can relax the constraints on the forms of regularization and data terms (*e.g.*, discrete optimization methods do not necessitate that the functions are differentiable) and is less susceptible to local minima compared to continuous optimization methods. Even though the global optimum cannot always be guaranteed, recent MRF optimization techniques provide a gap index to show how far the resulting energy is from the global optimum. Last but not least, as an important component of graph-based methods, graphical models combine probability theory and graph theory within a general formalism for modeling and solving inference problems using a Bayesian formulation. Such an approach has potential advantages in terms of parameter learning and uncertainty analysis over classic variational methods due to the probabilistic interpretation of the obtained solution [Szeliski 2010]. The aforementioned strengths of graph-based modeling and inference have resulted in the heavy adoption of these methods towards solving many computer vision, computer graphics and medical imaging problems. However, it is important to mention that the community has primarily focused on low-rank graphical models where interactions between parameters was often at the level of pair of variables. This was a convenient approach driven mostly from the optimization viewpoint since numerous efficient algorithms exist for solving pairwise MRFs. Such interactions to certain extent can cope with rather complex vision problems (segmentation, estimation, motion analysis and object tracking, disparity estimation from calibrated views, *etc.*), in particular when the viewpoint of the camera has little impact on the modeling process. However, in a number of visual perception tasks, either the camera pose or the "object" plays a fundamental role. This is often addressed through an alternating approach where given the pose parameters, inference on the graph is performed and the obtained solution is propagated back to the pose space towards re-estimating the pose.

Such a context motivated us to revise several of the most fundamental vision problems and to develop graph-based formulations for them. More specifically, the problems that we address in this thesis include:

- Joint multi-object tracking, segmentation and depth ordering from monocular 2D video sequences
- Non-rigid 3D surface matching
- Knowledge-based 3D model inference from 2D and/or 3D images

They are related to 2D, 2.5D, 3D and 2D-3D visions and thus can be regarded as a representative set of visual perception. Moreover, another motivation to solve these problems is originated from the interest of applications. Such fundamental problems are involved in numerous important vision applications, such as video surveillance, action recognition, robot navigation, shape/object recognition, deformation transfer, human-machine interaction and medical imaging.

However, "there is no free lunch", graph-based modeling is not straightforward for these vision problems due to the fact that the direct factorization of existing global objective functions is usually impossible. For example, depth ordering is usually expressed as a strict and total order between objects and thus involves all unknown variables, resulting in a challenging factorization requirement of the objective function defined on such an ordering. Similar difficulties are shared by 3D model inferences and surface matching problems which are often modeled using an objective function that strongly depends on both the global pose and local deformations. While global modeling is more intuitive and better studied, we have to resort to distributed models in order to achieve graph-based formulations.

1.1 Thesis Statement

In this thesis, we propose graph-based formulations for modeling the problems of interest stated above, so that various cues can be fused in a principled way and the variables of interest can be jointly inferred using discrete optimization techniques.

The overall methodology is to first investigate the global structure of each problem and then to search for proper local interactions the accumulation of which can globally constrain the configuration of the whole system. Since local interactions are encoded within local potential functions involving each a small number of variables, such interactions must be independent from the configuration of other variables in order to achieve a rigorous distributed model. Hence, the key step for graph-based modeling is to determine such "invariant" local interactions for each specific problem. Moreover, we usually expect the cardinality of local interactions (*i.e.*, the number of involved variables) to be as small as possible in order to inherit reasonable inference complexity.

Following such a methodology, we have developed a joint 2.5D layered model where top-down object-level and bottom-up pixel-level representations are seamlessly combined through local constraints involving only pairs of variables. Then, based on such a layered model, we have proposed for the first time a single-shot optimization framework for jointly performing segmentation, depth ordering and multi-object tracking from monocular video sequences using a pairwise MRF. Promising experimental results demonstrate the potential of this method and its robustness to noise, cluttered backgrounds, moving cameras and even complete occlusions.

For the problem of non-rigid 3D surface matching, we have developed a higher-order graph-based formulation that combines multiple measurements of geometric/appearance similarities and deformation prior. The use of higher-order interactions are motivated by the fact that three point correspondences between two surfaces can determine intrinsic deformation errors under the most natural assumption (*i.e.*, isometry) on the deformation between two surfaces. Through a number of challenging experiments, our approach was proved to robustly establish the correspondence between non-rigid surfaces undergoing large deformations, partial matching as well as inconsistent boundaries and scales.

Furthermore, we have used higher-order interactions to build a statistical shape model that is pose-invariant. Based on such a shape model, we have introduced a novel approach for knowledge-based 3D segmentation using a higher-order MRF, which does not require the estimation of the global pose or the initialization of the shape model. This approach has been validated on challenging data in the context of the human calf muscle segmentation.

Last but not least, for the problem of landmark-based 3D model inference from monocular 2D images, we have proposed a graph-based approach to simultaneously determine both the optimal 3D model and the corresponding 2D projections. To the best of our knowledge, this is the first attempt that can address both problems without explicit estimation of the camera viewpoint. We are in addition able to encode visibility modeling and therefore to deal with erroneous detections, lack of correspondences and/or partially visible configurations. Promising results on standard face benchmarks demonstrate the potential of our approach.

1.2 Outline of the Dissertation

The remainder of the dissertation is organized as follows. In chapter 2, we provide a survey on graphical models, which composes the background of the works presented in this thesis. Particular attention is given to the development of MRF models and their optimization techniques that are highly related to our methods. After that, the main works of this thesis are presented in chapters 3, 4 and 5, respectively. More specifically, in chapter 3, we present a novel joint layered model and a pairwise MRF formulation for simultaneously and jointly performing segmentation, multi-object tracking and depth ordering. A higher-order graph-based 3D surface matching method is introduced in chapter 4. In chapter 5, we propose one-shot optimization formulations for knowledge-based 3D segmentation and for 3D model inference from monocular 2D images, as well as a pose-invariant 3D shape prior. Finally, we conclude the thesis and discuss future works in chapter 6.

Chapter 2

Survey of Graphical Models

Graphical models combine probability theory and graph theory towards a natural and powerful formalism for modeling and solving inference and estimation problems in various scientific and engineering fields. They have several useful properties that one can benefit during the algorithm design:

1. A graph-based framework usually inherits *modularity*. Even though the whole system can be complex, the designs of different components are independent to some extent, and probability theory provides a principled way to combine these components together. Furthermore, the *modularity* also includes the fact that the modeling and the inference in such a framework are largely decoupled, which makes feasible the adoption of inference methods being developed in different fields.

2. The graph theoretic side of graphical models provides a simple way to visualize the structure of a model. Furthermore, these approaches encompass conditional independence properties, which facilitates the choice and design of parametric inference representations within the aforementioned context.

3. The factorization of the joint probability over a graph could produce inference problems that can be solved in a computational efficient manner. In particular, development of inference methods based on discrete optimization[1] enhances the potential of graphical models and enlarges significantly the set of visual perception problems on which the methods can be applied. Furthermore, the use of discrete optimization as inference methods can relax the constraints on the characteristics

[1] We should note that continuous graphical models have also been used in the literature (*e.g.*, [Isard 2003, Sigal *et al.* 2003, Sudderth *et al.* 2010]).

of the local functions (*e.g.*, discrete optimization methods do not necessitate that the functions are differentiable) and has better behavior in terms of convergence to a global minimum compared to continuous methods. Even though the global optimum cannot be guaranteed in general, state-of-the-art MRF optimization techniques (*e.g.*, TRW algorithms [Wainwright *et al.* 2005, Kolmogorov 2006] and dual-decomposition [Komodakis *et al.* 2007a]) provide a gap index to show how far the resulting energy is from the global optimum.

4. The variables in graphical models can be continuous and/or discrete, resulting in a better flexibility and capacity for the modeling with respect to other approaches such as *variational methods* [Tikhonov & Arsenin 1977, Engl *et al.* 1996].

5. The probabilistic side of graphical models leads to potential advantages in terms of parameter learning (*e.g.*, [Roth & Black 2007, Salakhutdinov 2009]) and uncertainty analysis (*e.g.*, [Kohli & Torr 2008, Glocker *et al.* 2008b]) over classic variational methods, due to the introduction of probability explanation to the solution [Szeliski 2010].

Hence, graphical models have been widely used in computer vision community, where problems (image restoration, image segmentation, stereo vision, *etc.*) often require to infer the latent states for a large number of variables of interest. In particular, *Undirected Graphical Models*, also known as *Markov Random Fields* (MRFs), have become a ubiquitous tool to model and solve vision problems.

This chapter provides a survey of graphical models, which is the cornerstone of our works presented in the following chapters of this dissertation. Our survey consists of two parts. The first part (section 2.1) introduces the three common types of graphical models, *i.e.*, directed graphical models, undirected graphical models and factor graphs. In particular, different subclasses of undirected graphical models are discussed as well as their applications in computer vision. The second part of this chapter (section 2.2) presents representative techniques for the MAP inference in discrete MRFs, where emphasis is paid on the methods that are closely related with the ones employed in this thesis.

2.1 Graphical models

A probabilistic graphical model consists of a graph where each node is associated with a random variable and an edge between a pair of nodes encodes probabilistic interaction between the corresponding variables. Each of such models provides a compact representation for a family of joint probability distributions which satisfy the conditional independence properties determined by the topology/structure of the graph: the associated family of joint probability distributions can be factorized into a product of local functions each of which involves a (usually small) subset of variables. Such a factorization is the key idea of graphical models.

There are two common types of graphical models: *Directed Graphical Models* (also known as *Bayesian Networks* or *Belief Networks*) and *Undirected Graphical Models* (also known as *Markov Random Fields* or *Markov Networks*), corresponding to directed and undirected graphs, respectively. They are used to model different families of distributions with different kinds of conditional independences. It is usually convenient to covert both of them into a unified representation which is called *Factor Graph*, in particular for performing inference. We will proceed with a formal brief presentation of each model where emphasis will be given to the ones which are strongly related with the content of this dissertation. We suggest the reader being interested for a larger and more in depth overview the following publications [Lauritzen 1996, Bishop 2006, Jordan 2007, Koller & Friedman 2009].

2.1.1 Preliminary Notations

Let us introduce the necessary notations that will be used throughout the dissertation.

For a graphical model, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the corresponding graph which consists of a set $\mathcal{V}$ of nodes and a set $\mathcal{E}$ of edges. Then, for each node i $(i \in \mathcal{V})$ contained in the model, let X_i denote the associated random variable, x_i the realization of X_i, and $\mathcal{X}_i$ the state space of x_i (*i.e.*, $x_i \in \mathcal{X}_i$). Also, let $\mathbf{X} = (X_i)_{i \in \mathcal{V}}$ denote the joint random variable and $\mathbf{x} = (x_i)_{i \in \mathcal{V}}$ the realization (configuration) of the graphical model taking values in its space $\mathcal{X}$ which is defined as the Cartesian product of the spaces for all individual variables, *i.e.*, $\mathcal{X} = \prod_{i \in \mathcal{V}} \mathcal{X}_i$.

For the purposes of simplification and concreteness, we use “probability distribution” to refer to “probability mass function” (with respect to the counting measure) in discrete cases and “probability density function” (with respect to the Lebesgue measure) in continuous cases. Furthermore, we use $p(x)$ to denote the probability distribution on a random variable X, and use x_c $(c \subseteq \mathcal{V})$ as the shorthand for a tuple c of variables, *i.e.*, $x_c = (x_i)_{i \in c}$. Due to the one-to-one mapping between a node and the associated random variable, for the purpose of convenience, we often use “node” to refer to the corresponding random variable in cases where there is no ambiguity.

2.1.2 Bayesian Networks (Directed Graphical Models)

A *Bayesian Network (BN)* has the structure of a directed acyclic graph (DAG) $\mathcal{G}$ where the edges in $\mathcal{E}$ are directed and no directed cycle exists (*e.g.*, Fig. 2.1(a)), and holds the following local independence assumptions (called *local Markov property*) which impose that every node is independent of its non-descendant nodes[2] given all its parents:

$$\forall\, i \in \mathcal{V}, X_i \perp X_{\mathcal{A}_i} | X_{\pi_i} \tag{2.1}$$

where $\mathcal{A}_i$ and π_i denotes the set of non-descendant nodes and the set of parents for a node i in the graph $\mathcal{G}$, respectively, and $X_i \perp X_j | X_k$ denotes the statement that X_i and X_j are independent given X_k. The associated family of joint probability distributions are those satisfying the local independences in Eq. 2.1, and can be factorized into the following form according to $\mathcal{G}$:

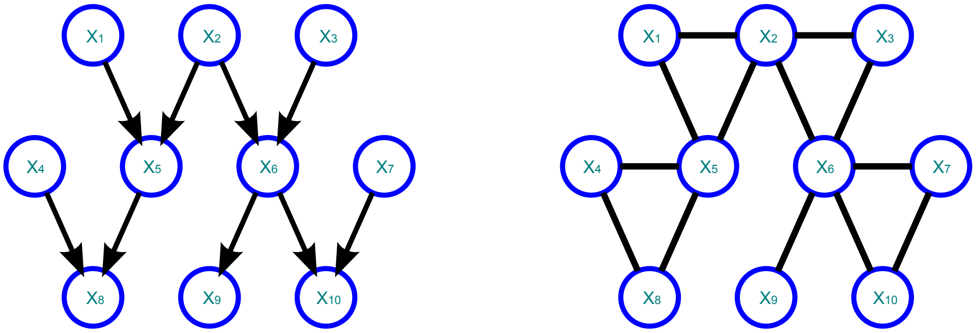
$$p(\mathbf{x}) = \prod_{i \in \mathcal{V}} p(x_i | x_{\pi_i}) \tag{2.2}$$

where $p(x_i|x_{\pi_i})$ denotes local conditional probability distribution (CPD) of x_i given the states x_{π_i} of the parents. It should be noted that any distribution with the factorized form in Eq. 2.2 satisfies the local independences in Eq. 2.1.

All conditional independences (called *global Markov property*) implied within the structure of BNs, including the local independences of Eq. 2.1, can be identified by checking *d-separation* properties of the corresponding graph $\mathcal{G}$ [Pearl 1988]. This can be performed using an intuitive and handy method: *Bayes ball algorithm* [Geiger *et al.* 1990, Shachter 1998]. Let $\mathcal{I}(\mathcal{G})$ denote the set of such conditional independences. Note that the global Markov property and the local Markov property are equivalent in BNs. Hence, if a distribution can be factorized over $\mathcal{G}$, it must satisfy all the conditional independences in $\mathcal{I}(\mathcal{G})$. On the other hand, we should also note that an instance of distribution that can be factorized over $\mathcal{G}$ may satisfy more independences than those in $\mathcal{I}(\mathcal{G})$. Nevertheless, such instances are very "few" in the sense that they have measure zero in the space of CPD parameterizations, *e.g.*, a slight perturbation of the local CPDs will almost certainly eliminate these "extra" independences [Koller & Friedman 2009].

BNs are usually used to model causal relationships between random variables and have been applied in many fields such as artificial intelligence, computer vision, automatic control, information engineering, *etc*. In computer vision, Hidden Markov Models (HMM) [Rabiner 1989] and Kalman Filters [Kalman 1960, Gelb 1974], which are well-known subsets of BNs, provide a common way to model temporal relations and has been employed to deal with object tracking [Terzopoulos & Szeliski 1993, Wu *et al.* 2002], denoising [Kim & Woods 1997, Romberg *et al.* 2001], motion analysis [Hervieu *et al.* 2007, Gui *et al.* 2008], sign language recognition [Starner *et al.* 1998, Moni & Ali 2009], *etc*. Besides, neural networks [Bishop 1995], another special type of BNs, provide an important machine learning method to deal with vision problems [Egmont-Petersen *et al.* 2002]. Other vision applications include for example [Pavlovic 1999] and [Zhang & Ji 2005],

[2] For a node $i \in \mathcal{V}$, its non-descendant nodes consist of the nodes $j \in \mathcal{V} - \{i\}$ such that there is no directed path from i to j.



(a) Bayesian Network

(b) Markov Random Filed

Figure 2.1: Examples of Bayesian Network and Markov Random Filed. Note that the directed graph in (a) can be transformed into the undirected graph in (b) by *moralization* process [Jordan 2007].

where dynamic BNs have been used to perform gesture/speech recognition and facial expression understanding, respectively.

2.1.3 Markov Random Fields (Undirected Graphical Models)

A *Markov Random Field (MRF)* has the structure of an undirected graph $\mathcal{G}$ where all edges of $\mathcal{E}$ are undirected (*e.g.*, Fig. 2.1(b)). Furthermore, such a paradigm inherits the following local independence assumptions (also called *local Markov property*):

$$\forall\, i \in \mathcal{V}, X_i \perp X_{\mathcal{V}-\{i\}} | X_{\mathcal{N}_i} \tag{2.3}$$

which impose that a node is independent of any other node given all its neighbors. In such a context, $\mathcal{N}_i = \{j | \{i, j\} \in \mathcal{E}\}$ denotes the set of neighbors of node i in the graph $\mathcal{G}$. An important notion in MRFs is *clique*, which is defined as a full-connected subset of nodes in the graph. A clique is *maximal* if it is not contained within any other larger clique. The associated family of joint probability distributions are those satisfying the local Markov property (*i.e.*, Eq. 2.3). According to Hammersley-Clifford theorem [Hammersley & Clifford 1971, Besag 1974], they are *Gibbs distributions* which can be factorized into the following form according to $\mathcal{G}$:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c) \tag{2.4}$$

where Z is the normalizing factor (also known as the partition function), $\psi_c(x_c)$ denotes the *potential function* of a clique c which is a positive real-valued function on the possible configuration x_c of the clique c, and $\mathcal{C}$ denotes a set of cliques[3] contained in the graph $\mathcal{G}$. We can also verify that any distribution with the factorized form in Eq. 2.4 satisfies the local Markov property in Eq. 2.3.

The *global Markov property* consists of all the conditional independences implied within the structure of MRFs, which are defined as: $\forall \mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3 \subseteq \mathcal{V}$, if any path from a node in $\mathcal{V}_1$ to a node in $\mathcal{V}_2$ includes at least one node in $\mathcal{V}_3$, then $X_{\mathcal{V}_1} \perp X_{\mathcal{V}_2} | X_{\mathcal{V}_3}$. Let $\mathcal{I}(\mathcal{G})$ denote the set of such conditional independences. The identification of these independences boils down to a "reachability" problem in graph theory: considering a graph $\mathcal{G}'$ which is obtained by removing the nodes in $\mathcal{V}_3$ as well as the edges connected to these nodes from $\mathcal{G}$, $X_{\mathcal{V}_1} \perp X_{\mathcal{V}_2} | X_{\mathcal{V}_3}$ is true if and only if there is no path in $\mathcal{G}'$ that connects any node in $\mathcal{V}_1 - \mathcal{V}_3$ and any node in $\mathcal{V}_2 - \mathcal{V}_3$. This problem can be solved using standard search algorithms such as breadth-first search (BFS) [Cormen *et al.* 2009]. Note that the local Markov property and the global Markov property are equivalent for any positive distribution. Hence, if a positive distribution can be factorized into the form in Eq. 2.4 according to $\mathcal{G}$, then it satisfies all the conditional independences in $\mathcal{I}(\mathcal{G})$. Similar to Bayesian Network, an instance of distribution that can be factorized over $\mathcal{G}$, may satisfies more independences than those in $\mathcal{I}(\mathcal{G})$.

MRFs provide a principled probabilistic framework to model vision problems, thanks to their ability to model soft contextual constraints between random variables [Li 2009]. The adoption of such constraints is important in vision problems, since the image and/or scene modeling involves interactions between a subset of pixels and/or scene components. Often, these constraints are referred to as "prior" of the whole system. Through MRFs, one can use nodes to model variables of interest and combine different available cues that can be encoded by clique potentials within a unified probabilistic formulation. Then the inference can be performed via *Maximum a posteriori* (MAP) estimation:

$$\mathbf{x}^{\text{opt}} = \arg\max_{\mathbf{x}} p(\mathbf{x}) \tag{2.5}$$

Since the potential functions are restricted to positive here, let us define clique energy θ_c as a real function on a clique c ($c \in \mathcal{C}$):

$$\theta_c(x_c) = -\log \psi_c(x_c) \tag{2.6}$$

[3] Note that any quantities defined on a non-maximal clique can always be redefined on the corresponding maximal clique, and thus $\mathcal{C}$ can also consist of only the maximal cliques. However, using only maximal clique potentials may obscure the structure of original cliques by fusing together the potentials defined on a number of non-maximal cliques into a larger clique potential. Compared with such a maximal representation, a non-maximal representation clarifies specific features of the factorization and usually leads to computational efficiency in practice. Hence, without loss of generality, we do not assume that $\mathcal{C}$ consist of only maximal cliques in this dissertation.

Due to the one-to-one mapping between θ_c and ψ_c, we also call θ_c *potential function* (or *clique potential*) on clique c in the remaining of this dissertation towards a more convenient representation of the joint distribution $p(\mathbf{x})$:

$$p(\mathbf{x}) = \frac{1}{Z} \exp\{-E(\mathbf{x})\} \tag{2.7}$$

where $E(\mathbf{x})$ denotes the *energy* of the MRF and is defined as a sum of potential functions on the cliques:

$$E(\mathbf{x}) = \sum_{c \in \mathcal{C}} \theta_c(x_c) \tag{2.8}$$

Since the "-log" transformation between the distribution $p(\mathbf{x})$ and the energy $E(\mathbf{x})$ is a monotonic function, the MAP inference in MRFs (*i.e.*, the maximization of $p(\mathbf{x})$ in Eq. 2.5) is equivalent to the minimization of $E(\mathbf{x})$ as follows:

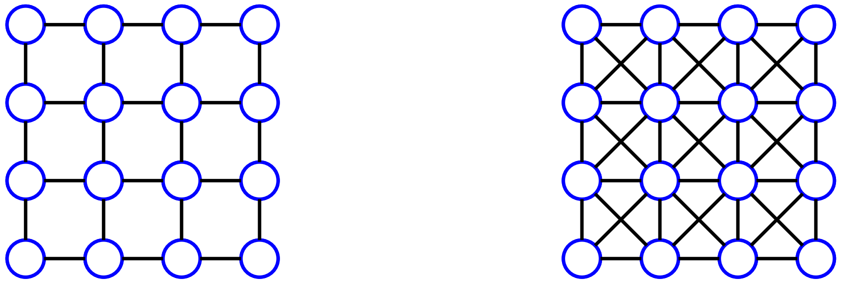
$$\mathbf{x}^{\text{opt}} = \arg\min_{\mathbf{x}} E(\mathbf{x}) \tag{2.9}$$

In cases of *discrete MRFs* where the random variables are discrete (*i.e.*, $\forall\, i \in \mathcal{V}$, $\mathcal{X}_i$ consists of a discrete label set), the above optimization becomes a discrete optimization problem. Numerous works have been done to develop efficient MRF optimization/inference algorithms using discrete optimization theories and techniques (*e.g.*, [Boykov *et al.* 2001, Ishikawa 2003, Kolmogorov & Zabih 2004, Wainwright *et al.* 2005, Kohli & Torr 2007, Kolmogorov 2006, Komodakis *et al.* 2008, Pawan Kumar *et al.* 2009, Komodakis 2010]), which have been successfully employed to efficiently solve vision problems using MRF-based methods (*e.g.*, [Kolmogorov & Zabih 2002, Glocker *et al.* 2008a, Kohli *et al.* 2008b, Szeliski *et al.* 2008, Boykov & Funka-Lea 2006]). We will provide a survey on an important subset of such works in section 2.2. Due to the advantages regarding both the modeling and the inference as discussed above, discrete MRFs have been widely employed to solve vision problems. Below, we present several typical subsets of MRFs commonly used in vision community.

Pairwise MRF Models

The most common type of MRFs that is widely used in computer vision is the *pairwise MRF*, in which the associated energy is factorized into a sum of potential functions defined on cliques of order strictly less than three. More specifically, a pairwise MRF consists of a graph $\mathcal{G}$ with a set $(\theta_i(\cdot))_{i\in\mathcal{V}}$ of *singleton potentials* (also known as *unary potentials*) defined on single variables and a set $(\theta_{ij}(\cdot))_{\{i,j\}\in\mathcal{E}}$ of *pairwise potentials* defined on pairs of variables. The MRF energy has the following form:

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}} \theta_i(x_i) + \sum_{\{i,j\}\in\mathcal{E}} \theta_{ij}(x_{ij}) \tag{2.10}$$



(a) 4-neighborhood system

(b) 8-neighborhood system

Figure 2.2: Examples of MRFs with Grid-like Structures

Pairwise MRFs have attracted the attention of a lot of researchers and numerous works have been done in past decades, mainly due to the facts that pairwise MRFs inherit simplicity and computational efficiency. On top of that, their use was spread due to the fact that the interaction between pairs of variables is the most common and fundamental type of interactions required to model many vision problems. In computer vision, such works include both the modeling of vision problems using pairwise MRFs (*e.g.*, [Geman & Geman 1984, Rother *et al.* 2004, Felzenszwalb & Huttenlocher 2005, Boykov & Funka-Lea 2006]) and the efficient inference in pairwise MRFs (*e.g.*, [Boykov *et al.* 2001, Wainwright *et al.* 2005, Kolmogorov 2006, Kohli & Torr 2007, Komodakis *et al.* 2007a]). Two of the most important graph structures used in computer vision are *grid-like structures* (*e.g.*, Fig. 2.2) and *pictorial structures* (*e.g.*, Fig. 2.3). Grid-like structures provide a natural and reasonable representation for images, while pictorial structures are often associated with deformable (articulated) objects.

Pairwise MRFs of *grid-like structures* (*e.g.*, Fig. 2.2) have been widely used in computer vision to deal with a large number of important problems, such as image denoising/restoration (*e.g.*, [Geman & Geman 1984, Greig *et al.* 1989]), stereo vision/multi-view reconstruction (*e.g.*, [Roy & Cox 1998, Kolmogorov & Zabih 2002, Vogiatzis *et al.* 2007]), optical flow and motion analysis (*e.g.*, [Black & Anandan 1993, Sun *et al.* 2010]), image registration and matching (*e.g.*, [Glocker *et al.* 2008a, Shekhovtsov *et al.* 2008]), segmentation (*e.g.*, [Boykov & Kolmogorov 2003, Rother *et al.* 2004, Boykov & Funka-Lea 2006]) and over-segmentation (*e.g.*, [Moore *et al.* 2010, Veksler *et al.* 2010]).

In this context, the nodes of an MRF correspond to the lattice of pixels[4] and the edges corresponding to pairs of neighbor nodes are considered to encode contextual constraints between nodes. The random variable x_i associated with each node i represents a physical

[4] Other homogeneously distributed unit such as control point [Glocker *et al.* 2008a] can also be considered in such MRFs.

quantity specific to problems (*e.g.*, an index denoting the segment that the corresponding pixel belongs to for image segmentation problem, an integral value between 0 and 255 denoting the intensity of the corresponding pixel for gray image denoising problem, *etc.*). The data likelihood is encoded by the sum of the singleton potentials $\theta_i(\cdot)$, whose definition is specific to the considered applications (*e.g.*, for image denoising, such singleton terms are often defined as a penalty function based on the deviation of the observed value from the underlying value.). The contextual constraints compose a prior model on the configuration of the MRF, which is usually encoded by the sum of all the pairwise potentials $\theta_{ij}(\cdot,\cdot)$. The most typical and commonly used contextual constraint is the *smoothness*, which imposes that physical quantities corresponding to the states of nodes varies "smoothly" in the spatial domain as defined by the connectivity of the graph. To this end, the pairwise potential $\theta_{ij}(\cdot,\cdot)$ between a pair $\{i,j\}$ of neighbor nodes is defined as a cost term that penalizes the variation of the states between the two nodes:

$$\theta_{ij}(x_{ij}) = \rho(x_i - x_j) \tag{2.11}$$

where $\rho(\cdot)$ is usually an even and non-decreasing function. In computer vision, common choices (Eq. 2.12) for $\rho(\cdot)$ are *(generalized) Potts model*[5] [Potts 1952, Boykov *et al.* 1998], *truncated absolute distance* and *truncated quadratic*, which are typical *discontinuity preserving* penalties:

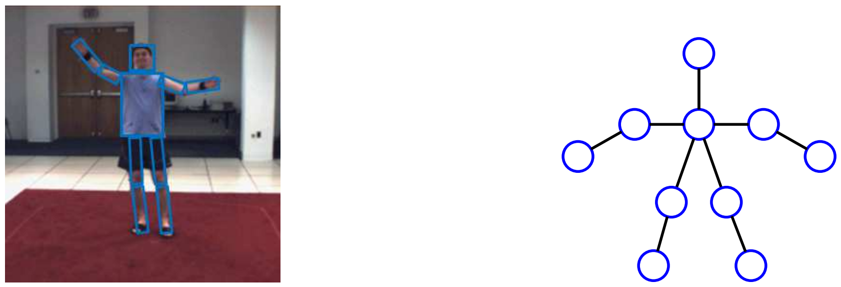
$$\rho(x_i - x_j) = \begin{cases} w_{ij}\cdot(1-\delta(x_i - x_j)) & \text{(Potts models)} \\ \min(K_{ij}, w_{ij}\cdot|x_i - x_j|) & \text{(truncated absolute distance)} \\ \min(K_{ij}, w_{ij}\cdot(x_i - x_j)^2) & \text{(truncated quadratic)} \end{cases} \tag{2.12}$$

where $w_{ij} \geq 0$ is a weight coefficient[6] for the penalities, Kronecker delta $\delta(x)$ is equal to 1 when $x = 0$ and 0 otherwise, and K_{ij} is a coefficient representing the maximum penalty allowed in the truncated models. More discontinuity preserving regularization functions can be found in for example [Terzopoulos 1986, Lee & Pavlidis 1988]. Such discontinuity preserving terms reduce the risk of over-smoothing, which is an advantage compared with Total Variation (TV) regularizations [Chan & Shen 2005] that are often used in variational methods [Tikhonov & Arsenin 1977, Engl *et al.* 1996].

MRFs of *pictorial structures* (*e.g.*, Fig. 2.3) provide a powerful part-based modeling tool for representing deformable objects and in particular articulated objects. Their nodes correspond to components of such objects. The corresponding latent variables represent the spatial pose of the components. An edge between a pair of nodes encode the interactions such as kinematic constraints between the corresponding pair of components. In

[5] Note that *Ising model* [Ising 1925, Geman & Geman 1984] is a particular case of *Potts model* where each node has two possible states.

[6] w_{ij} is a constant for all pairs $\{i,j\}$ of nodes in the original Potts model in [Potts 1952].



(a) Pictorial Model (b) MRF corresponding to the Pictorial Model in (a)

Figure 2.3: Example of MRFs with Pictorial Structures (The original image used in (a) is from *HumanEva-I* database: http://vision.cs.brown.edu/humaneva/.)

[Felzenszwalb & Huttenlocher 2005], *Pictorial model* [Fischler & Elschlager 1973] was introduced into computer vision to deal with pose recognition of human body and face. In this work, a tree-like MRF (see Fig. 2.3) was employed to model the spring-like prior between pairs of components through pairwise potentials, while the data likelihood is encoded in the singleton potentials each of which is computed from the appearance model of the corresponding component. The pose parameters of all the components are estimated though the MAP inference, which can be done very efficiently in such a tree-structured MRF using dynamic programming [Bellman 1957, Cormen *et al.* 2009] (*i.e.*, min-sum belief propagation [Pearl 1988, Yedidia *et al.* 2003, Bishop 2006]). This work has gained a lot of attention in computer vision and the proposed part-based models have been adopted and/or extended to deal with the pose estimation, detection and tracking of deformable object such as human body [Sigal *et al.* 2003, Sigal & Black 2006a, Eichner & Ferrari 2009, Andriluka *et al.* 2009], hand [Sudderth *et al.* 2004b, Sudderth *et al.* 2004a] and other objects [Pawan Kumar *et al.* 2004, Felzenszwalb *et al.* 2010]. In [Pawan Kumar *et al.* 2004], part-based model of [Felzenszwalb & Huttenlocher 2005] was extended regarding the topology of the MRF as well as the image likelihood in order to deal with the pose estimation of animals such as cows and horses. Continuous MRFs of pictorial structures were proposed in [Sigal *et al.* 2003] and [Sudderth *et al.* 2004b] to deal with body and/or hand tracking, where nonparametric belief propagation algorithms [Isard 2003, Sudderth *et al.* 2010] were employed to perform inference. In the subsequent papers [Sigal & Black 2006a, Sudderth *et al.* 2004a], occlusion reasoning was introduced into their graphical models in order to deal with occlusions between different components. Indeed, the wide existence of such occlusions in the cases of articulated objects is an important limitation of the part-based modeling. The modeling of occlusions in graphical models is still an open problem.

Higher-order MRF Models

Higher-order MRFs (also known as *high-order MRFs*) involve potential functions that are defined on cliques containing more than two nodes and cannot be further decomposed. One can express conveniently these graphical models by grouping the cliques according to their order:

$$E(\mathbf{x}) = \sum_{k=1}^{K} \sum_{c \in \mathcal{C}_k} \theta_c(x_c) \tag{2.13}$$

where $\mathcal{C}_k$ denotes the set of cliques of order k and K denotes the highest order in the model.

Higher-order MRFs are often used to model more complex and/or natural statistics between random variables and richer interactions between them. One can cite for example the higher-order MRF model proposed in [Roth & Black 2005, Roth & Black 2009] to better characterize image priors, by using the Product-of-Experts framework to define the higher-order potentials. Such a higher-order model was successfully applied in image denoising and inpainting problems [Roth & Black 2005, Roth & Black 2009]. $\mathcal{P}^n$ *Potts model* was proposed in [Kohli *et al.* 2007, Kohli *et al.* 2009b], which consists of a strict generalization of the generalized Potts model [Boykov *et al.* 1998] (see Eq. 2.12). It considers a similar interaction between n nodes (instead of between two nodes) and its performance was demonstrated in image segmentation being a natural application domain of such a model. In [Kohli *et al.* 2008a, Kohli *et al.* 2009a], $\mathcal{P}^n$ Potts model was further enriched towards a *robust* $\mathcal{P}^n$ *model*, which produced better segmentation performance. Higher-order smoothness priors were used in [Woodford *et al.* 2009] to solve stereo reconstruction problems. Other types of higher-order pattern potentials were also considered in [Komodakis & Paragios 2009] to deal with image/signal denoising and image segmentation problems. All these works demonstrated that the inclusion of higher-order interactions is able to improve the performance compared to pairwise models in the considered vision problems.

Higher-order models become even more important in the cases where we need to model measures that intrinsically involve more than two variables. A simple example is the modeling of second-order derivative (or even higher-order derivatives), which is often used to measure bending force in shape prior modeling such as active contour models (*i.e.*, "Snake") [Kass *et al.* 1988]. In [Amini *et al.* 1990], dynamic programming was adopted to solve "Snake" model in a discrete setting, which is essentially a higher-order MRF model. A third-order spatial priors based on second derivatives was also introduced to deal with image registration in [Kwon *et al.* 2008]. In the optical flow formulation proposed in [Glocker *et al.* 2010], higher-order potentials were used to encode angle deviation prior, non-affine motion prior as well as the data likelihood.

More recently, global models, which include potentials involving all the nodes, have been developed, together with the inference algorithms for them. One can cite for example [Vicente *et al.* 2008] and [Nowozin & Lampert 2009] where global connectivity priors (*e.g.*, foreground segment must be connected) were used to enforce the connectedness of the resulting labels for binary image segmentation, [Delong *et al.* 2010] where 'label costs" [Zhu & Yuille 1996] was introduced into graph-based segmentation formulation to deal with unsupervised image segmentation, and [Ladicky *et al.* 2010a, Ladicky *et al.* 2011] which proposed to incorporate "object co-occurrence statistics" in Conditional Random Field (CRF) models to object class image segmentation.

Conditional Random Fields

A Conditional Random Field (CRF) [Lafferty *et al.* 2001, Sutton & McCallum 2011] encodes, with the same concept as the MRF earlier described, a conditional distribution $p(\mathbf{X}|\mathbf{D})$ where $\mathbf{X}$ denotes a tuple of latent variables and $\mathbf{D}$ a tuple of observed variables (data). It can be viewed as an MRF which is globally conditioned on the observed data $\mathbf{D}$. Accordingly, the Markov properties for the CRF are defined on the conditional distribution $p(\mathbf{X}|\mathbf{D})$. The local Markov properties in such a context become:

$$\forall\, i \in \mathcal{V}, X_i \perp X_{\mathcal{V}-\{i\}} | \{X_{\mathcal{N}_i}, \mathbf{D}\} \tag{2.14}$$

while the global Markov property can also be defined accordingly. The conditional distribution $p(\mathbf{X}|\mathbf{D})$ over the latent variables $\mathbf{X}$ is also a Gibbs distribution and can be written as the following form:

$$p(\mathbf{x}|\mathbf{D}) = \frac{1}{Z(\mathbf{D})} \exp\{-E(\mathbf{x};\mathbf{D})\} \tag{2.15}$$

where the energy $E(\mathbf{x};\mathbf{D})$ of the CRF is defined as:

$$E(\mathbf{x};\mathbf{D}) = \sum_{c \in \mathcal{C}} \theta_c(x_c;\mathbf{D}) \tag{2.16}$$

We can observe that there is no modeling on the probabilistic distribution over the variable in $\mathbf{D}$, which relaxes the concern on the dependencies between these observed variables, whereas such dependencies can be rather complex. Hence, CRFs reduce significantly difficulty in modeling the joint distribution of the latent and observed variables, and observed variables can be incorporated into the CRF framework in a more flexible way. Such a flexibility is one of the most important advantages of CRFs compared with generative MRFs[7] when used to model a system. For example, the fact that clique potentials can be data dependent in CRFs could lead to more informative interactions than data independent

[7] Like [Pawan Kumar 2008], we use the term *generative MRFs* to distinguish the usual MRFs from CRFs.

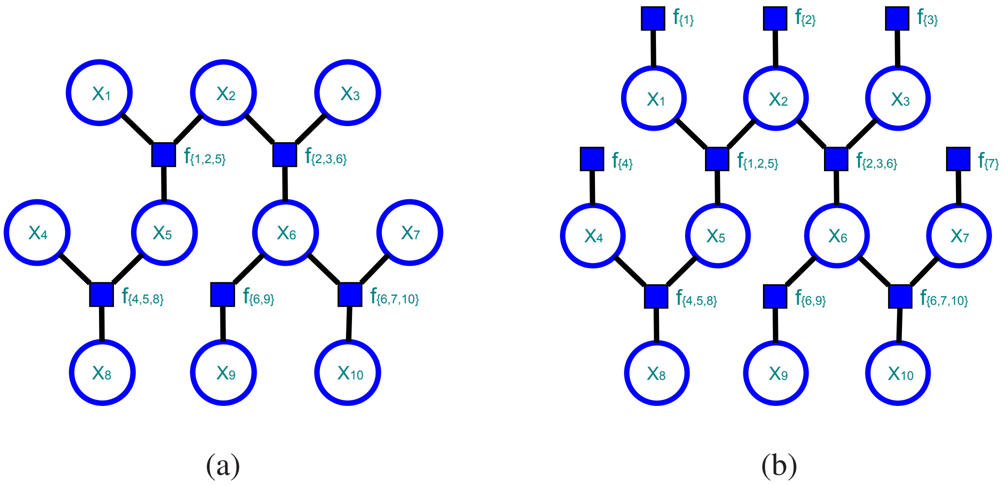


Figure 2.4: Examples of Factor Graphs. Note that both of the Bayesian Network in Fig. 2.1(a) and the Markov Random Filed in Fig. 2.1(b) can be represented by the two factor graphs above. However, the factor graph in (b) contains factors corresponding to non-maximal cliques.

clique potentials. Such an concept was adopted for example in binary image segmentation [Boykov & Jolly 2001].

CRFs have been applied to various fields such as computer vision, bioinformatics and text processing among others. In computer vision, for example, grid-like CRFs was introduced in [Kumar & Hebert 2004] to model spatial dependencies in the image, an approach that outperformed the classic MRF model [Geman & Geman 1984] in the image restoration experiments. A multi-scale CRF model was proposed in [He *et al.* 2004] for object class image segmentation, and a more sophisticated model named “associative hierarchical CRFs” were proposed in [Ladicky *et al.* 2009] to solve the same problem. Following that, in [Ladicky *et al.* 2010b], object detectors and CRFs were combined within a CRF model which can be solved efficiently, so as to jointly estimate the class category, location, and segmentation of objects/regions from 2D images. CRFs has been also applied for object recognition. For example, a discriminative part-based approach was proposed in [Quattoni *et al.* 2004] to recognize objects based on a tree-structured CRF.

Despite the difference in the probabilistic explanation, the MAP inferences in generative MRFs and CRFs boil down to the same problem. For the purpose of convenience, we do not explicitly represent the observed variables in the graph in this dissertation, however, the implied model (generative MRFs or CRFs) will be clear in the context.

2.1.4 Factor Graphs

Factor graph [Frey 1998, Kschischang *et al.* 2001] is a unified representation for both BNs and MRFs, which uses additional nodes, named *factor nodes*[8], to explicitly describe the factorization of the joint distribution in the graph.

More specifically, a set $\mathcal{F}$ of factor nodes are introduced into the graph, corresponding each to an objective function term defined on a subset of usual nodes. Each factor encodes a local conditional probability distribution defined on a usual node and its parents in cases of BNs (see Eq. 2.2), while it encodes a potential function defined on a clique in cases of MRFs (see Eq. 2.4 or Eq. 2.8). The associated joint probability is a product of factors:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{f \in \mathcal{F}} \phi_f(x_f) \tag{2.17}$$

where the normalizing factor Z is equal to 1 for BNs. Similar to MRFs, we can define the energy of the factor graph as:

$$E(\mathbf{x}) = \sum_{f \in \mathcal{F}} \theta_f(x_f) \tag{2.18}$$

where $\theta_f(x_f) = -\log \phi_f(x_f)$. Note that there can be more than one factor graphs corresponding to a BN or MRF. Fig. 2.4 shows two examples of factor graphs which provide two different possible representations for both the Bayesian Network in Fig. 2.1(a) and the Markov Random Filed in Fig. 2.1(b).

Factor graphs are bipartite, since there are two types of nodes and no edge exists between two nodes of same types. Such a representation conceptualizes in a clear manner the underlying factorization of the distribution in the graphical model. In particular for MRFs, factor graphs provide a feasible representation to describe explicitly the cliques and the corresponding potential functions when non-maximal cliques are also considered (*e.g.*, Fig. 2.4(b)). The same objective can be hardly met using the usual graphical representation of MRFs. Computational inference is another strength of factor graphs representations. The *sum-product* and *min-sum* (or: *max-product*[9]) algorithms in the factor graph [Kschischang *et al.* 2001, Bishop 2006] generalize the classic counterparts [Pearl 1988, Yedidia *et al.* 2003] in the sense that the order of factors can be greater than two, which will be presented in section 2.2.2. Furthermore, since an MRF with loops may has no loop in its corresponding factor graph (*e.g.*, see the MRF in Fig. 2.1(b) and the factor graphs in Fig. 2.4 (a-b)), in such cases the *min-sum* algorithm in the factor graph can

[8] We call the nodes in original graphs *usual nodes* when an explicit distinction between the two types of nodes is required to avoid ambiguities.

[9] The *max-product* algorithm is to maximize the probability $p(\mathbf{x})$ which is a product of local functions (Eq. 2.17), while the *min-sum* algorithm is to minimize the corresponding energy which is a sum of local energy functions (Eq. 2.18). They are essentially the same algorithm.

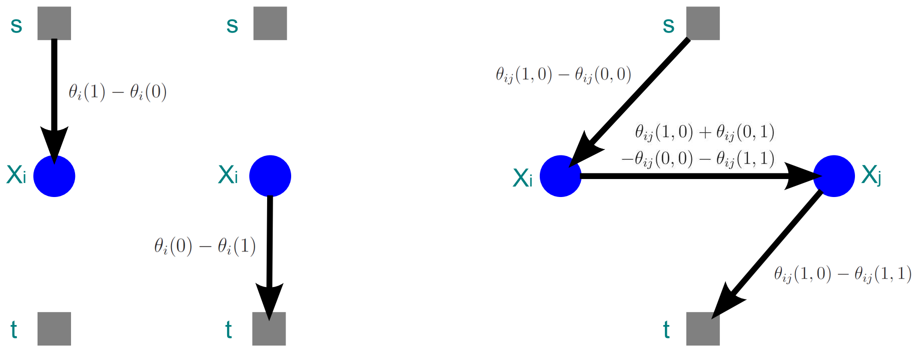
perform the MAP inference exactly with polynomial complexity. Let us call such factor graphs without loop (*e.g.*, Fig. 2.4 (a-b)) as *Factor tree*. Such an important subset of factor graphs will be used later in this dissertation.

2.2 MAP Inference Methods for Discrete MRFs

An essential problem regarding the application of MRF models is how to infer the value for each of the nodes contained in an MRF. This thesis focuses on the MAP inference (*i.e.*, Eq. 2.5) in discrete MRFs, which boils down to an energy minimization problem as shown in Eq. 2.9. Such a combinatorial problem is known to be NP-hard in general [Boykov *et al.* 2001, Kolmogorov & Zabih 2004], except for some particular cases such as MRFs of bounded tree-width [Dawid 1992, Aji & McEliece 2000, Jordan 2007] (*e.g.*, tree-structured MRFs [Pearl 1988]) and pairwise MRFs with submodular energy [Kolmogorov & Zabih 2004, Schlesinger & Flach 2006].

The most well-known early (before early 1990s) algorithms for optimizing the MRF energy were *iterated conditional modes* (ICM) [Besag 1986], *simulated annealing* methods (*e.g.*, [Geman & Geman 1984, Blake & Zisserman 1987, Tupin *et al.* 1998]) and *highest confidence first (HCF)* [Chou & Brown 1990, Chou *et al.* 1993]. While being computational efficient methods, ICM and HCF suffer from their ability to recover a good optimum. On the other hand, for simulated annealing methods, even if in theory they provide certain guarantees on the quality of the obtained solution, in practice from computational viewpoint such methods are impractical. In the 1990s, more advanced methods, such as *loopy belief propagation* (LBP) [Freeman *et al.* 2000, Weiss & Freeman 2001, Felzenszwalb & Huttenlocher 2006] and *graph cuts* techniques (*e.g.*, [Greig *et al.* 1989, Roy & Cox 1998, Boykov *et al.* 1998, Ishikawa & Geiger 1998, Boykov *et al.* 2001]), provided powerful alternatives to the aforementioned methods from both computational and theoretical viewpoint and have been used to solve numerous visual perception problems (*e.g.*, [Freeman *et al.* 2000, Sun *et al.* 2003, Greig *et al.* 1989, Ishikawa & Geiger 1998, Kolmogorov & Zabih 2002, Boykov & Kolmogorov 2003, Rother *et al.* 2004]). Since then, the MRF optimization is experiencing a renaissance, and more and more researchers have been working on it. For the most recent MRF optimization techniques, one can cite for example *QPBO* techniques [Boros *et al.* 1991, Kolmogorov & Rother 2007, Boros *et al.* 2006, Rother *et al.* 2007], LP primal-dual algorithms (*e.g.*, [Komodakis *et al.* 2008]) as well as dual methods (*e.g.*, [Wainwright *et al.* 2005, Kolmogorov 2006, Komodakis *et al.* 2007b, Werner 2007]). All these advances in the MRF optimization make the application of MRFs more and more popular and become a ubiquitous tool in computer vision.

A brief overview of inference methods that are often employed in computer vision community, in particular the techniques that have been adopted in the works of this thesis,


s
$\theta_i(1) - \theta_i(0)$
Xi
t
s
Xi
$\theta_i(0) - \theta_i(1)$
t
s
$\theta_{ij}(1,0) - \theta_{ij}(0,0)$
$\theta_{ij}(1,0) + \theta_{ij}(0,1)$
$-\theta_{ij}(0,0) - \theta_{ij}(1,1)$
Xi
Xj
$\theta_{ij}(1,0) - \theta_{ij}(1,1)$
t

(a) (b)

Figure 2.5: Examples of *s-t Graph* Construction for Binary Graph Cuts [Kolmogorov & Zabih 2004]. (a) Graphs for the singleton potential defined on a node i. The left one is for the cases where $\theta_i(0) < \theta_i(1)$ and the right one is for the cases where $\theta_i(0) \geq \theta_i(1)$; (b) Graph for the pairwise potential defined on an edge $\{i, j\}$ where $\theta_{ij}(1,0) > \theta_{ij}(0,0)$ and $\theta_{ij}(1,0) > \theta_{ij}(1,1)$. Note that $\theta_{ij}(1,0) + \theta_{ij}(0,1) - \theta_{ij}(0,0) - \theta_{ij}(1,1) > 0$ holds when the energy is submodular.

will be presented in the upcoming sections. To this end, we will first review binary *Graph cuts* and their extensions for minimizing the energy of pairwise MRFs in section 2.2.1. Then in section 2.2.2, we will describe the *min-sum* belief propagation algorithm in factor tree and also show its extensions towards dealing with an arbitrary graphical model. Following that, we review in section 2.2.3 recent developed dual methods for pairwise MRFs, in particular the *tree-reweighted message passing* methods (*e.g.*, [Wainwright *et al.* 2005, Kolmogorov 2006]) and the *dual-decomposition* approaches (*e.g.*, [Komodakis *et al.* 2007b, Komodakis *et al.* 2011]). Last but not least, a survey on inference methods for higher-order MRFs will be provided in section 2.2.4.

2.2.1 Graph Cuts and Extensions

Graph cuts consist of a family of discrete algorithms that use *min-cut/max-flow* techniques to efficiently minimize the energy of discrete MRFs and have been used to solve various vision problems (*e.g.*, [Greig *et al.* 1989, Ishikawa & Geiger 1998, Rother *et al.* 2004, Kolmogorov & Zabih 2002, Boykov & Funka-Lea 2006, Kohli *et al.* 2008b]).

The basic idea of graph cuts is to construct a directed graph $\mathcal{G}^{st} = (\mathcal{V}^{st}, \mathcal{E}^{st})$ (called *s-t graph*[10], see examples in Fig. 2.5) with two special terminal nodes (*i.e.*, the source s and the sink t) and non-negative capacity setting $c(i, j)$ on each directed edge $(i, j) \in \mathcal{E}^{st}$, such

[10] Note that generations such as *multi-way cut* problem [Dahlhaus *et al.* 1992] which involves more than

that the cost $C(S,T)$ (Eq. 2.19)) of the s-t cut that partitions the nodes into two disjoint sets (S and T such that $s \in S$ and $t \in T$) is equal[11] to the energy of the MRF with the corresponding configuration[12] $\mathbf{x}$.

$$C(S,T) = \sum_{i \in S, j \in T, (i,j) \in \mathcal{E}^{st}} c(i,j) \tag{2.19}$$

An MRF that has such an s-t graph is called *graph-representable*[13] and can be solved in polynomial time using graph cuts [Kolmogorov & Zabih 2004]. The minimization of the energy of such an MRF is equivalent to the minimization of the cost of the s-t-cut problem (*i.e.*, min-cut problem). The Ford and Fulkerson theorem [Ford & Fulkerson 1962] states that the solution of the min-cut problem corresponds to the maximum flow from the source s to the sink t (*i.e.*, max-flow problem). Such a problem can be efficiently solved in polynomial time using many existing algorithms such as Ford-Fulkerson style augmenting paths algorithms [Ford & Fulkerson 1962] and Goldberg-Tarjan style push-relabel algorithms [Goldberg & Tarjan 1988]. Note that the min-cut problem and the max-flow problem are actually dual LP problems of each other [Vazirani 2001].

Unfortunately, not all the MRFs are graph-representable. Previous works have been done to explore the class of graph-representable MRFs (*e.g.*, [Boros & Hammer 2002, Ishikawa 2003, Kolmogorov & Zabih 2004, Schlesinger & Flach 2006]) and demonstrated that a pairwise discrete MRF is graph-representable so that the global minimum of the energy can be achieved in polynomial time via *Graph cuts* if the energy function of the MRF is *submodular*. There are various definitions of *submodular* energy functions in the literature that are equivalent. We consider here the one used in [Schlesinger & Flach 2006]. Let us assume $\mathcal{X}_i$ ($\forall\, i \in \mathcal{V}$) to be a completely ordered set, the energy function of a pairwise discrete MRF is *submodular* if each pairwise potential term θ_{ij} ($\forall\, \{i,j\} \in \mathcal{E}$) satisfies: $\forall\, x_i^1, x_i^2 \in \mathcal{X}_i$ *s.t.* $x_i^1 \leq x_i^2$, and $\forall\, x_j^1, x_j^2 \in \mathcal{X}_j$ *s.t.* $x_j^1 \leq x_j^2$,

$$\theta_{ij}(x_i^1, x_j^1) + \theta_{ij}(x_i^2, x_j^2) \leq \theta_{ij}(x_i^1, x_j^2) + \theta_{ij}(x_i^2, x_j^1), \tag{2.20}$$

For binary cases where the $\mathcal{X}_i = \{0,1\}$ ($\forall\, i \in \mathcal{V}$), the condition is reduced to that each pairwise potential θ_{ij} ($\forall\, \{i,j\} \in \mathcal{E}$) satisfy:

$$\theta_{ij}(0,0) + \theta_{ij}(1,1) \leq \theta_{ij}(0,1) + \theta_{ij}(1,0) \tag{2.21}$$

two terminal nodes are NP-hard.

[11] There may be a constant difference between the cost of cut and the MRF energy.

[12] The following rule can be used to associate an s-t cut to an MRF labeling: for a node $i \in \mathcal{V}^{st} - \{s,t\}$, i) if $i \in S$, the label x_i of the corresponding node in the MRF is equal to 0; ii) if $i \in T$, the label x_i of the corresponding node in the MRF is equal to 1.

[13] Note that, in general, such an s-t graph is not unique for a graph-representable MRF.

However, in numerous vision problems, more challenging energy functions are often required that do not satisfy the submodular condition in Eq. 2.20. The minimization of such non-submodular energy functions are NP-hard in general [Boykov *et al.* 2001, Kolmogorov & Zabih 2004] and an approximation algorithm would be required to approach the global optimum.

In vision community, [Greig *et al.* 1989] proposed to use min-cut/max-flow techniques to exactly optimize the energy of a binary (*i.e.*, binary-label) MRF (Ising model) for image restoration in polynomial time. However, such techniques did not draw much attention in vision in the following decade since then, probably due to the fact that the model considered in [Greig *et al.* 1989] is quite simple. Such a situation has changed in late 1990s when a number of techniques based on Graph cuts were proposed to solve more complicated MRFs (*e.g.*, multi-labels MRFs). One can cite for example the works of [Roy & Cox 1998], [Boykov *et al.* 1998] and [Ishikawa & Geiger 1998], which proposed to use min-cut/max-flow techniques to minimize multi-label MRFs.

Since then, numerous works have been done for exploring larger subsets of MRFs that can be exactly or approximately optimized by graph cuts and for developing more efficient graph cuts algorithms. We can cite for example an efficient graph construction method proposed in [Ishikawa 2003] to deal with arbitrary *convex* pairwise MRFs. In [Boykov *et al.* 2001], α-*expansion* and $\alpha\beta$-*swap* were introduced to generalize binary Graph cuts to handle pairwise MRFs with *metric* and/or *semi-metric* energy with optimum quality guarantee (*i.e.*, the ratio between the obtain energy and the global optimal energy is bounded by a factor). An important problem was studied in [Kolmogorov & Zabih 2004], *i.e.*, what kinds of MRF energy functions can be minimized by Graph cuts. Besides, [Kolmogorov & Zabih 2004] also introduced a more efficient graph construction approach compared to [Boykov *et al.* 2001] and proposed a method able to deal with the minimization of third-order pseudo-boolean functions. A dynamic max-flow algorithm was proposed in [Kohli & Torr 2005, Kohli & Torr 2007] to accelerate graph cuts when dealing with dynamics MRFs (*i.e.*, the potential functions vary over time, whereas the change between two successive instants is usually quite small), where the key idea is to reuse the flow obtained by solving the previous MRF so as to significantly reduce the computational time of min-cut. Another dynamic algorithm was also proposed in [Juan & Boykov 2006] to improve the convergence of optimization for dynamic MRFs, by using the min-cut solution of the previous MRF to generate an initialization for solving the current MRF. In [Komodakis *et al.* 2007b] and [Komodakis *et al.* 2008], a primal-dual scheme based on linear programming relaxation was proposed for optimizing the MRF energy. This method can be viewed as a generalization of α-expansion and achieves a substantial speedup with respect to previous methods such as [Boykov *et al.* 2001] and [Komodakis & Tziritas 2007]. Two similar but simpler techniques with respect to that of [Komodakis *et al.* 2007b] were proposed in [Alahari *et al.* 2008] to achieve a similar computational efficiency. Besides, an efficient algorithm based on max-flow and elimination techniques was introduced in [Carr & Hartley 2009] for the optimization of 4-neighborhood grid-like MRFs.

We should note that several methods do also exist for partially inferring solutions for non-submodular binary energy functions. About three decades ago, *Roof duality* was proposed in [Hammer *et al.* 1984], which provides a way to achieve a partial optimal labeling for quadratic pseudo-boolean functions (the solution will be a complete labeling that corresponds to global optimum if the energy is submodular). Such a method was efficiently implemented in [Boros *et al.* 1991], which is referred to as *Quadratic Pseudo-Boolean Optimization (QPBO)* algorithm and can be regarded as a graph-cuts-based algorithm with a special graph construction where two nodes in s-t graph are used to represent two complementary states of a node in the original MRF [Kolmogorov & Rother 2007]. By solving min-cut/max-flow in such an s-t graph, QPBO outputs a solution assigning 0, 1 or $\frac{1}{2}$ to each node in the original MRF, where the label $\frac{1}{2}$ means the corresponding node is *unlabeled*. Later, two different techniques were introduced in order to extend QPBO towards achieving a complete solution. One is *probing* (called *QPBO-P*) [Boros *et al.* 2006, Rother *et al.* 2007], which aims to gradually reduce the number of unlabeled nodes (either by finding the optimal label for certain unlabeled nodes or by regrouping a set of unlabeled nodes) until convergence by iteratively fixing the label of a unlabeled node and performing QPBO. The other one is *improving* (called *QPBO-I*) [Rother *et al.* 2007], which starts from a complete labeling $\mathbf{y}$ and gradually improves such a labeling by iteratively fixing the labels of a subset of nodes as those specified $\mathbf{y}$ and using QPBO to get a partial labeling to update $\mathbf{y}$. These QPBO techniques were further extended in [Kohli *et al.* 2008c] to deal with multi-label MRFs, where the key idea is to convert a multi-label MRF into an equivalent binary MRF [Ishikawa 2003] and then use QPBO techniques to solve the linear relaxation of the obtained binary MRF. For the inference in multi-label MRFs, another interesting method based on QPBO and move techniques was proposed in [Lempitsky *et al.* 2010], which is referred to as *fusion moves*. Different from previous move techniques such as α-expansion and $\alpha\beta$-swap, such a method fuses two arbitrary proposals of the full labeling by using QPBO and achieves a new labeling that has an energy less or equal than the energies of both proposals.

2.2.2 Belief Propagation Algorithms

Belief propagation algorithms use local message passing to perform inference on graphical models. These methods provide an exact inference algorithm for tree-structured graphical models, while an approximate solution can be achieved when a loopy graph is considered. For those loopy graphs with low tree-widths (Eq. 2.23) such as cycles, extended belief propagation methods such as *junction tree algorithm* [Dawid 1992, Aji & McEliece 2000, Jordan 2007] provide an efficient algorithm to perform exact inference.

Belief Propagation in Tree

Belief propagation (BP) [Pearl 1988, Yedidia *et al.* 2003, Bishop 2006] was proposed originally for exactly solving MAP inference (*min-sum* algorithm) and/or maximum-marginal inference (*sum-product* algorithm) in a tree-structured graphical model in polynomial time. These methods can be viewed as a special case of *dynamic programming* in graphical models [Bellman 1957, Cormen *et al.* 2009, Felzenszwalb & Zabih 2011].

The *min-sum* algorithm[14] is described in Algorithm 2.2 using the factor graph representation [Kschischang *et al.* 2001, Bishop 2006], since as we mentioned in section 2.1.4, the factor graph makes the BP algorithm applicable to more cases compared to the classic min-sum algorithm applied on a usual pairwise graph [Freeman *et al.* 2000]. In general, the complexity of the belief propagation in the tree is $O(NL^K)$, where N, L, K denote the number of nodes, the number of candidates for each node, and the maximum order of the factors, respectively. Note that *reparameterization* (also known as *equivalent transformation*) of the MRF energy (*e.g.*, [Wainwright *et al.* 2004, Kolmogorov 2006]) provides an alternative interpretation of belief propagation and leads to a memory-efficient implementation [Kolmogorov 2006].

Loopy Belief Propagation

The tree-structured constraint limits the use of the standard belief propagation algorithm presented above. In computer vision, most of the problems require loopy graphical models to encode well the interactions between variables. Hence, researchers have investigated to extend the message passing concept for minimization of arbitrary graphs.

Loopy belief propagation (LBP), a natural step towards this direction, performs message passing iteratively in the graph (*e.g.*, [Frey & MacKay 1998, Freeman *et al.* 2000, Weiss & Freeman 2001, Felzenszwalb & Huttenlocher 2006]) despite of the existence of loops. We refer the reader to [Freeman *et al.* 2000, Weiss & Freeman 2001] for the details and discussion on the LBP algorithm. Regarding the message passing scheme in loopy graphs, there are two possible choices: *parallel* or *sequential*. In the parallel scheme, messages are computed for all the edges at the same time and then the messages are propagated for the next round of message passing. Whereas in the sequential scheme, a node

[14] Note that all the BP-based algorithms presented in section 2.2.2 include both *min-sum* and *sum-product* versions. We focus here on the *min-sum* version, since we consider MAP inference in the works that have been done in this thesis. Nevertheless, one can easily obtain the *sum-product* version by replacing the message computation with the sum of the product of function terms. We refer the reader to [Kschischang *et al.* 2001, Bishop 2006, Jordan 2007] for more details.

propagates the message to one of its neighbor node at each round and such a message will be used to compute the messages sent by that neighbor node. [Tappen & Freeman 2003] showed empirically that the sequential scheme was significantly faster than the parallel one, while the performance of both methods was almost the same. Substantial investment was made towards improving the efficiency of message passing by exploiting different types of structure regarding the graph and/or the potential functions. For example, an efficient method was proposed in [Pawan Kumar & Torr 2006] to reduce computational and memory cost for *robust truncated models* where a pairwise potential is equal to a constant for most of the state combination of the two nodes. [Felzenszwalb & Huttenlocher 2006] introduced a strategy for speeding up belief propagation for cases where pairwise potential functions only depend on the difference of the variables such as those defined in Eq. 2.12, an approach to accelerating the message passing in bipartite graphs (including grid-like MRFs in Fig. 2.2), and a multi-scale belief propagation scheme to perform inference in grid-like MRFs. Two speed-up techniques specifically for grid-like MRF models were also proposed in [Petersen *et al.* 2008].

Despite the fact that LBP performed well for a number of vision applications such as [Freeman *et al.* 2000, Sun *et al.* 2003], they cannot guarantee to converge to a fixed point, while their theoretical properties are not well understood. Last but not least, their solution is generally worse than more sophisticated generalizations of message passing algorithms (*e.g.*, [Wainwright *et al.* 2005, Kolmogorov 2006, Komodakis *et al.* 2007a]) that will be presented in section 2.2.3 [Szeliski *et al.* 2008].

Algorithm 2.1 Ordering of the Nodes for Sending Messages In a Tree

```
Input: Tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V}$ and edge set $\mathcal{E}$
Input: Root node $\hat{r} \in \mathcal{V}$
Output: $\mathcal{P}_{\text{send}} = \text{NodeOrdering}(\mathcal{T}, \hat{r})$, where $\mathcal{P}_{\text{send}}$ is a list denoting the ordering of the
  nodes in tree $\mathcal{T}$ for sending messages
  $\mathcal{P}_{\text{send}} \leftarrow (\hat{r})$
  if $|\mathcal{V}| > 1$ then
    Get the set $\mathcal{C}$ of child nodes: $\mathcal{C} \leftarrow \{i | i \in \mathcal{V}, \{i, \hat{r}\} \in \mathcal{E}\}$
    for all $c \in \mathcal{C}$ do
      Get child tree $\mathcal{T}_c$ with root $c$
      $\mathcal{P}_{\text{send}} \leftarrow (\text{NodeOrdering}(\mathcal{T}, \hat{r}), \mathcal{P}_{\text{send}})$   $\{\mathcal{P}_{\text{send}}$ is ordered from left to right$\}$
    end for
  end if
  return $\mathcal{P}_{\text{send}}$
```

Algorithm 2.2 Min-sum Belief Propagation in Factor Tree

```
Input: Factor tree $\mathcal{T} = (\mathcal{V} \cup \mathcal{F}, \mathcal{E})$ with usual node set $\mathcal{V}$, factor node set $\mathcal{F}$ and edge set $\mathcal{E}$
Input: Factor potentials $(\theta_f(\cdot))_{f \in \mathcal{F}}$
Output: The optimal configuration $\mathbf{x}^{\text{opt}} = \arg\min_{\mathbf{x}} \sum_{f \in \mathcal{F}} \theta_f(x_f)$
  Choose a node $\hat{r} \in \mathcal{V}$ as the root of the tree
  Construct $\Pi$ s.t. $\Pi(i)$ denotes the parent of node $i \in \mathcal{V} \cup \mathcal{F}$
  Construct $\boldsymbol{\mathcal{C}}$ s.t. $\boldsymbol{\mathcal{C}}(i)$ denotes the set of children of node $i \in \mathcal{V} \cup \mathcal{F}$
  $\mathcal{P}_{\text{send}} \leftarrow \text{NodeOrdering}(\mathcal{T}, \hat{r})$ {see Algorithm 2.1}
  for $k = 1 \rightarrow \text{length}(\mathcal{P}_{\text{send}}) - 1$ do
    $i \leftarrow \mathcal{P}_{\text{send}}(k)$
    parent node $p \leftarrow \Pi(i)$
    child node set $\mathcal{C} \leftarrow \boldsymbol{\mathcal{C}}(i)$
    if $i \in \mathcal{V}$ then
      if $|\mathcal{C}| > 0$ then
        $m_{i \rightarrow p}(x_i) \leftarrow \sum_{j \in \mathcal{C}} m_{j \rightarrow i}(x_i)$
      else
        $m_{i \rightarrow p}(x_i) \leftarrow 0$
      end if
    else
      if $|\mathcal{C}| > 0$ then
        $m_{i \rightarrow p}(x_p) \leftarrow \min_{x_{\mathcal{C}}} (\phi(x_i) + \sum_{j \in \mathcal{C}} m_{j \rightarrow i}(x_j))$
        $s_i(x_p) \leftarrow \arg\min_{x_{\mathcal{C}}} (\phi(x_i) + \sum_{j \in \mathcal{C}} m_{j \rightarrow i}(x_j))$
      else
        $m_{i \rightarrow p}(x_p) \leftarrow \phi(x_p)$ {$p$ is the unique variable contained in factor $i$ in this case.}
      end if
    end if
  end for
  $x_{\hat{r}}^{\text{opt}} \leftarrow \arg\min_{x_{\hat{r}}} \sum_{j \in \boldsymbol{\mathcal{C}}(\hat{r})} m_{j \rightarrow \hat{r}}(x_{\hat{r}})$
  for $k = \text{length}(\mathcal{P}_{\text{send}}) - 1 \rightarrow 1$ do
    $i \leftarrow \mathcal{P}_{\text{send}}(k)$
    if $i \in \mathcal{F}$ then
      parent node $p \leftarrow \Pi(i)$
      child node set $\mathcal{C} \leftarrow \boldsymbol{\mathcal{C}}(i)$
      $x_{\mathcal{C}}^{\text{opt}} \leftarrow s_i(x_p)$
    end if
  end for
  return $\mathbf{x}^{\text{opt}}$
```

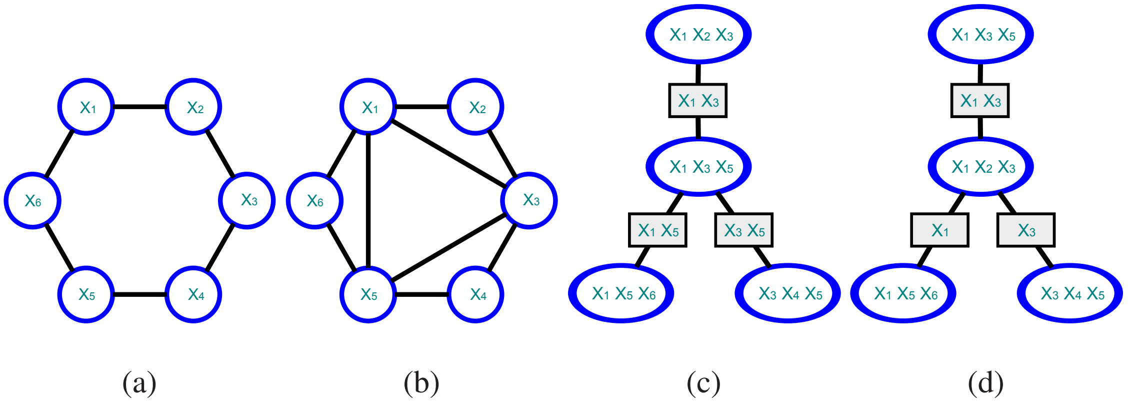


Figure 2.6: Example of Junction Tree. (a) Original undirected graphical model; (b) Triangulation of the graph in (a); (c) A junction tree for the graphs in (a) and (b); (d) A clique tree which is not junction tree.

Junction Tree Algorithm

Junction tree algorithm (JTA) is an exact inference method in arbitrary graphical models [Dawid 1992, Aji & McEliece 2000, Jordan 2007]. The key idea is to make systematic use of the Markov properties implied in graphical models to decompose a computation of the joint probability or energy into a set of local computations. Such an approach bears strong similarities with message passing in the standard belief propagation or dynamic programming. In this sense, we regard JTA as an extension of the standard belief propagation. Let us introduce some necessary notions and properties about junction trees and then discuss briefly the corresponding inference algorithm.

For a clique set $\mathcal{C}$, the corresponding *clique tree* is defined as a tree-structured graph $\mathcal{G}_J$ with node set $\mathcal{V}_J$ and edge set $\mathcal{E}_J$ where each node i ($i \in \mathcal{V}_J$) represents a clique $c_i \in \mathcal{C}$. A *junction tree* is a clique tree which processes the junction tree property: for every pair of cliques c_i and c_j in $\mathcal{G}_J$, $c_i \cap c_j$ is contained in all the cliques on the (unique) path between c_i and c_j. The junction tree property ensures that local consistency implies global consistency so that local message passing process can produce exact inference. The example in Fig. 2.6 provides two clique trees (Fig. 2.6(c) and (d)) corresponding to the undirected graph in Fig. 2.6(b), where we use square boxes to explicitly represent the separators each of which is associated to an edge and denotes the intersection of the two cliques connected by the edge. We can easily verify that the clique tree in Fig. 2.6(c) is a *junction tree*, while the other one in Fig. 2.6(d) is not.

There are two important properties about junction trees [Jordan 2007], which are useful for the construction of a junction tree given an undirected graphical model:

1. An undirected graph has a *junction tree* if and only if it is triangulated (*i.e.*, there is no *chordless*[15] cycle in the graph.

2. A clique tree is a junction tree if and only if it is a maximal spanning tree which is a clique tree that has the maximal weight (*i.e.*, $\sum_{i,j \in \mathcal{E}_J} |c_i \cap c_j|$) over all possible trees connecting the considered cliques.

Hence, for a given undirected graph (*e.g.*, Fig. 2.6(a)), we can first triangulate[16] it (*e.g.*, Fig. 2.6(b)), and then find a maximal spanning tree to form a junction tree for the maximal cliques contained in this triangulated graph. This operation will produce a junction tree for the undirected graph (*e.g.*, Fig. 2.6(c)). For each clique c in the original graph, the associated clique potential θ_c is accumulated to the potential $\hat{\theta}_i$ of one and only one node i in the junction tree such that c is included in the clique c_i corresponding to node i (*i.e.*, $c \subseteq c_i$).

Without considering optimality of the generated junction tree[17], the triangulation can be done easily using *undirected graph elimination* algorithm [Jordan 2007]. This method successively eliminates the nodes in a graph by connecting the remaining neighbors of the node and removing the node as well as the edge connected to it from the graph. The second step, *i.e.*, the finding of a maximal spanning tree, can be easily performed using greedy algorithms such as Kruskal's algorithm [Cormen *et al.* 2009].

The energy[18] of a junction tree is defined as a sum of the potentials of the cliques corresponding to the nodes:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}_J} \hat{\theta}_i(x_{c_i}) \tag{2.22}$$

where c_i denotes the clique corresponding to node i of the junction tree. Due to the junction tree property, we can perform local message passing in the junction tree to do the inference, which is similar to standard belief propagation in factor trees. Interestingly, nodes in junction trees can be regarded as factor nodes in factor trees, while separators in junction trees can be regarded as usual nodes (may corresponding to a set of variables) in factor trees. Then the belief propagation scheme in the junction tree can be obtained easily from the one for the factor tree (see Algorithm 2.2). Hence, we do not present the message passing process here to avoid redundancy and refer the reader to [Aji & McEliece 2000, Jordan 2007] for details.

[15] A cycle is said to be *chordless* if there is no edge between two nodes that are not successors in the cycle.

[16] For directed graphical models, a *moralization* process [Jordan 2007] is to be applied prior to the triangulation in order to transform the directed graph to an undirected graph.

[17] Note that there may exist several such junction trees corresponding to an undirected graph. As we will discuss below, the optimality of a junction tree is related to its *width*. However, it is generally an NP-hard problem to find an optimal junction tree [Jordan 2007].

[18] The joint probability of a junction tree is defined as a product of potential functions, which is similar to factor graph in Eq. 2.17. We do not present it here for the purpose of compactness.

It can be easily noticed that in discrete cases, the complexity of the inference (*i.e.*, belief propagation) in a junction tree is exponential with respect to its *width* W. The width is defined as the maximum cardinal of the corresponding cliques over all nodes minus 1, *i.e.*:

$$W = \max_{i \in \mathcal{V}_J} |c_i| - 1 \tag{2.23}$$

Hence, the complexity is dominated by the largest maximal cliques in the triangulated graph. However, the triangulation process may produce large maximal cliques, while finding of an optimal junction tree with the smallest width for an arbitrary undirected graph is an NP-hard problem. Furthermore, graphical models with dense initial connections could lead to maximal cliques of very high cardinal even if an optimal junction tree could be found [Jordan 2007]. Due to the computational complexity, the junction tree algorithm becomes impractical when the tree width is high, although it provides an exact inference approach. Thus it has been only used in some specific scenarios or some special kinds of graphs that have low tree widths (*e.g.*, cycles and outer-planar graphs whose widths are equal to 2). For example, JTA was employed in [Paskin 2003] to deal with simultaneous localization and mapping (SLAM) problem, and was also adopted in [Batra *et al.* 2010] to perform exactly inference in outer-planar graphs within the whole dual-decomposition framework. In order to reduce the complexity, *nested junction tree* technique was proposed in [Kjæ rulff 1998] to further factorize large cliques. Nevertheless, the gain of such a process depends directly on the initial graph structure and is still insufficient to make JTA widely applicable in practice.

2.2.3 Dual Methods

The MAP inference in pairwise MRFs (Eq. 2.9, 2.10), can be reformulated as the *integer linear programming (ILP)* [Wainwright & Jordan 2007] as follows:

$$\begin{aligned} \min_{\boldsymbol{\tau}} \quad & E(\boldsymbol{\theta}, \boldsymbol{\tau}) = \langle \boldsymbol{\theta}, \boldsymbol{\tau} \rangle = \sum_{i \in \mathcal{V}} \sum_{a \in \mathcal{X}_i} \theta_{i;a} \tau_{i;a} + \sum_{(i,j) \in \mathcal{E}} \sum_{(a,b) \in \mathcal{X}_i \times \mathcal{X}_j} \theta_{ij;ab} \tau_{ij;ab} \\ s.t. \quad & \boldsymbol{\tau} \in \boldsymbol{\tau}^{\mathcal{G}} = \left\{ \boldsymbol{\tau} \;\middle|\; \begin{array}{ll} \sum_{a \in \mathcal{X}_i} \tau_{i;a} = 1 & \forall\, i \in \mathcal{V} \\ \sum_{a \in \mathcal{X}_i} \tau_{ij;ab} = \tau_{j;b} & \forall\, \{i,j\} \in \mathcal{E}, b \in \mathcal{X}_j \\ \tau_{i;a} \in \{0,1\} & \forall\, i \in \mathcal{V}, a \in \mathcal{X}_i \\ \tau_{ij;ab} \in \{0,1\} & \forall\, \{i,j\} \in \mathcal{E}, (a,b) \in \mathcal{X}_i \times \mathcal{X}_j \end{array} \right\}. \end{aligned} \tag{2.24}$$

where $\theta_{i;a} = \theta_i(a)$, $\theta_{ij;ab} = \theta_{ij}(a,b)$, binary variables[19] $\tau_{i;a} = [x_u = a]$ and $\tau_{ij;ab} = [x_i = a, x_j = b]$, $\boldsymbol{\tau}$ denotes the concatenation of all these binary variables which can be defined as $((\tau_{i;a})_{i \in \mathcal{V}, a \in \mathcal{X}_i}, (\tau_{ij;ab})_{\{i,j\} \in \mathcal{E}, (a,b) \in \mathcal{X}_i \times \mathcal{X}_j})$, and $\boldsymbol{\tau}^{\mathcal{G}}$ denotes the domain of $\boldsymbol{\tau}$.

[19] $[\cdot]$ is equal to one if the argument is true and zero otherwise.

Unfortunately the above ILP problem is NP-hard in general. Numerous approximation algorithms of MRF optimization have been developed based on *Linear Programming (LP)* relaxation of such a problem in Eq. 2.24, aiming to minimize $E(\boldsymbol{\theta}, \boldsymbol{\tau})$ in a relaxed domain $\hat{\boldsymbol{\tau}}^{\mathcal{G}}$ (called *local marginal polytope*) which is obtained by replacing the integer constraints in Eq. 2.24 by the non-negative constraints, *i.e.*:

$$\begin{aligned} \min_{\boldsymbol{\tau}} \quad & E(\boldsymbol{\theta}, \boldsymbol{\tau}) = \langle \boldsymbol{\theta}, \boldsymbol{\tau} \rangle = \sum_{i \in \mathcal{V}} \sum_{a \in \mathcal{X}_i} \theta_{i;a} \tau_{i;a} + \sum_{(i,j) \in \mathcal{E}} \sum_{(a,b) \in \mathcal{X}_i \times \mathcal{X}_j} \theta_{ij;ab} \tau_{ij;ab} \\ s.t. \quad & \boldsymbol{\tau} \in \hat{\boldsymbol{\tau}}^{\mathcal{G}} = \left\{ \boldsymbol{\tau} \,\middle|\, \begin{array}{ll} \sum_{a \in \mathcal{X}_i} \tau_{i;a} = 1 & \forall\, i \in \mathcal{V} \\ \sum_{a \in \mathcal{X}_i} \tau_{ij;ab} = \tau_{j;b} & \forall\, \{i,j\} \in \mathcal{E}, b \in \mathcal{X}_j \\ \tau_{i;a} \geq 0 & \forall\, i \in \mathcal{V}, a \in \mathcal{X}_i \\ \tau_{ij;ab} \geq 0 & \forall\, \{i,j\} \in \mathcal{E}, (a,b) \in \mathcal{X}_i \times \mathcal{X}_j \end{array} \right\}. \end{aligned} \tag{2.25}$$

For purposes of clarity, from now on, the term *MRF-MAP* will be used for the original MAP inference problem (Eq. 2.24) and *MRF-LP* for the relaxed one (Eq. 2.25).

It is generally infeasible to directly apply generic LP algorithms such as *interior point methods* [Boyd & Vandenberghe 2004] to solve MRF-LP problems corresponding to MRF models in computer vision [Yanover *et al.* 2006], due to the fact that the number of variables involved in $\boldsymbol{\tau}$ is usually huge. Instead, many methods in the literature have been designed based on solving some *dual* to the MRF-LP problem in Eq. 2.25, *i.e.*, maximizing the lower bound of $E(\boldsymbol{\theta}, \boldsymbol{\tau})$ provided by the dual. One can cite for example the *min-sum diffusion* [Kovalevsky & Koval 1975] and *augmenting DAG* [Koval & Schlesinger 1976] algorithms that were reviewed in [Werner 2007], the *message passing* algorithm based on *block coordinate descent* proposed in [Globerson & Jaakkola 2007], *tree-reweighted Message Passing (TRW)* techniques [Wainwright *et al.* 2005, Kolmogorov 2006] and *dual decomposition (MRF-DD)* [Komodakis *et al.* 2007b, Komodakis *et al.* 2011]. The tightening of the LP-relaxation has also been investigated towards achieving a better optimum of the MRF-MAP problem (*e.g.*, [Sontag & Jaakkola 2007, Komodakis & Paragios 2008, Pawan Kumar *et al.* 2009, Werner 2010]). Here, we review briefly the TRW and MRF-DD techniques, which have been used in the context of this thesis.

Tree-reweighted Message Passing

Tree-reweighted max-product message passing (TRW) algorithms [Wainwright *et al.* 2005, Kolmogorov 2006] are well-explored MRF optimization methods. The key idea of TRW algorithms is to solve the MRF-LP problem via a dual problem based on convex combination of trees. Actually, the optimal values of such a dual problem and of the MRF-LP problem coincide, since strong duality holds [Wainwright *et al.* 2005]. Furthermore, in

TRW algorithms, the LP relaxation (Eq. 2.25) is tight if a fix point of TRW algorithms satisfies a condition referred to as (strong) *tree agreement (TA)* [Wainwright *et al.* 2005], where a global optimal solution to the original MRF problem is achieved.

In [Wainwright *et al.* 2005], such an methodology was introduced to solve the MRF-MAP problem by using two different (edge-based and tree-based) message passing schemes, called *TRW-E* and *TRW-T*, respectively. These variants can be viewed as combinations of reparameterization and averaging operations on the MRF energy. However, both of the schemes do not guarantee the convergence of the algorithms and the value of the lower bound may fall into a loop. A sequential message passing scheme was proposed in [Kolmogorov 2006], which is known as *TRW-S*. Different from TRW-E and TRW-T, the TRW-S algorithm updates messages in a sequential order instead of a parallel order. Such a difference introduce to the algorithm better convergence properties, *i.e.*, the lower bound will not decrease. TRW-S will attain a point that satisfies a condition referred to as *weak tree agreement (WTA)* [Kolmogorov & Wainwright 2005] and the lower bound will not change any more since then[20]. Although the global optimum of the dual problem satisfies WTA condition, the converse is not necessarily true and therefore TRW-S cannot guarantee the global maximum of the lower bound in general. Nevertheless, as demonstrated in [Kolmogorov & Wainwright 2005], a WTA fixed point for the cases of binary pairwise MRFs always corresponds to the global maximum of the dual problem, and thus also corresponds to the global optimum of the MRF-LP problem. Furthermore, if a binary pairwise MRF is submodular, a WTA fixed point always achieves the global optimum of the MRF-MAP problem.

Dual Decomposition

In [Komodakis *et al.* 2007a, Komodakis *et al.* 2011], *dual-decomposition* [Bertsekas 1999] principle was introduced into the MRF optimization problem. The outcome was a general and powerful framework to minimize the MRF energy, which will be called *MRF-DD* in the remaining part of the thesis. The key idea of MRF-DD is: instead of minimizing directly the energy of the original problem (referred to as *master* problem) that is too complex to solve directly, we decompose the master problem into a set of subproblems (referred to as *slave* problems). The main characteristic of these subproblems is that each of them is easier to solve both in terms of cardinality as well as in terms of convexity. Once such decomposition is achieved, the solution of the master problem is obtained by combining the solutions of the slaves problems. Such an idea can be summarized mathematically as following: based on a Lagrangian dual of the *MRF-MAP* problem in Eq. 2.24,

[20] [Kolmogorov 2006] observed in the experiments that TRW-S would finally converge to a fixed point but such a convergence required a lot of time after attaining WTA. Nevertheless, such a convergence may not be necessary in practice, since the lower bound will not change any more after attaining WTA.

the sum of the minima of the slave problems that are obtained by the decomposition of the master problem provides a *lower bound* on the energy of the original MRF. This sum is maximized using *projected subgradient* method so that a solution to the master problem can be extracted from the Lagrangian solutions[21].

Such a MRF optimization framework possesses a great flexibility, generality and convergence property:

1. The Lagrangian dual problem can be globally optimized due to the convexity of the dual function. The solution obtained by the MRF-DD algorithm satisfies *weak tree agreement (WTA)* condition[22], while a solution satisfying WTA condition is not necessarily the optimum to the Lagrangian dual. The properties of *tree agreement* and *weak tree agreement* fix points [Kolmogorov & Wainwright 2005] are also applicable within the MRF-DD method.

2. Different decompositions of the master problem can be considered to deal with MRF-MAP problem. Each of such decompositions leads to a certain relaxation of the MRF-MAP problem. Interestingly, when the master problem is decomposed into a set of trees, the Lagrangian relaxation employed by MRF-DD is equivalent to the LP relaxation in Eq. 2.25, which is exactly the problem TRW algorithms aim to solve[23]. However, within MRF-DD framework, one can consider more sophisticated decompositions to tighten the relaxation (*e.g.*, decompositions based on outerplanar graphs [Batra *et al.* 2010] and K-fan graphs [Kappes *et al.* 2010]). To this end, a very useful theoretical conclusion has been drawn in [Komodakis *et al.* 2011] which provides an approach to comparing the tightness between two different decompositions.

3. Only MAP inference in slave problems are required and there is no constraints on how such an inference is done. As a result, one can apply specific optimization algorithms to solve slave problems and even different optimization algorithms for different slave problems. The natural outcome of such a property is high flexibility for designing new graph-based optimization algorithms based on such a dual decomposition framework. A number of elegant applications have been proposed in the literature, which include the graph matching method proposed in [Torresani *et al.* 2008],

[21] [Komodakis *et al.* 2011] provides a detailed discussion on different approaches to obtaining a feasible solution of the master problem from the solution of the slave problems after solving the Lagrangian dual.

[22] WTA condition can be easily extended to the cases where one or more slave problems are not tree-structured.

[23] The main difference between MRF-DD and TRW algorithms consists in the mechanism of the update of dual variables. The former relies on the optimal solution of slave problems while the latter is based on the min-marginals of the trees corresponding to slave problems.

the higher-order MRF inference method developed in [Komodakis & Paragios 2009], and the algorithm for joint segmentation and appearance histogram models optimization introduced in [Vicente *et al.* 2009].

However, computational cost is the main drawback of the MRF-DD algorithm. Reducing the running time for the convergence is an open problem and there are various techniques that have been proposed in the literature. For example, two approaches were proposed in [Komodakis 2010] to speed-up LP-based algorithms. One is to use a multi-resolution hierarchy of dual relaxations, and the other consists of a decimation strategy that gradually fixes the labels for a growing subset of nodes as well as their dual variables during the process. [Jojic *et al.* 2010] proposed to construct a smooth approximation of the energy function of the master problem by smoothing the energies of the slave problems so as to achieve a significant acceleration of the MRF-DD algorithm. A distributed implementation of graph cuts was introduced in [Strandmark & Kahl 2010] to solve the slave problems in parallel.

2.2.4 Inference in Higher-order MRFs

Higher-order potentials allow a better characterization of statistics between random variables and increase largely the ability of graph-based modeling. The rapid development of computer hardwares in terms of memory capacity and CPU speed also motivates the use of higher-order models in computer vision community. Nevertheless, efficient inference algorithms for solving higher-order MRFs are necessary towards expanding their use in vision problems that usually involve a large number of variables. In such a context, numerous works have been devoted in the past decade to search for inference algorithms in higher-order models. One can cite for example the work of Roth and Black [Roth & Black 2005, Roth & Black 2009], where a simple inference scheme based on a conjugate gradient method was developed to solve their higher-order model for image restoration. Since then, besides a number of methods for solving specific types of higher-order models (*e.g.*, [Kohli *et al.* 2007, Ramalingam *et al.* 2008, Nowozin & Lampert 2009, Delong *et al.* 2010, Ladicky *et al.* 2010a]), various techniques also have been proposed to deal with more general MRF models (*e.g.*, [Lan *et al.* 2006, Potetz & Lee 2008, Komodakis & Paragios 2009, Ishikawa 2009]). These inference methods are highly inspired from the ones for pairwise MRFs. Thus, similar to pairwise MRFs, there are also three main types of approaches for solving higher-order MRFs, *i.e.*, algorithms based on *reduction* and *graph cuts*, higher-order extensions of *belief propagation*, and *dual methods*.

Reduction* and *Graph cuts

Most of existing methods tackle inference in higher-order MRFs using a two-stage approach: first to reduce a higher-order model to a pairwise one with the same minimum, and then to apply standard methods such as graph cuts to solve the obtained pairwise model. The idea of order reduction exists for long time. More than thirty years ago, a method (referred to as *variable substitution*) was proposed in [Rosenberg 1975] to perform order reduction for models of any order, by introducing auxiliary variables to substitute products of variables[24]. However, this approach leads to a large number of non-submodular components in the resulting pairwise model. This is due to the hard constraints involved in the substitution, which causes large difficulty in solving the obtained pairwise model. This may explain why its impact is rather limited in the literature [Boros & Hammer 2002, Ali *et al.* 2008], since our final interest is solving higher-order models. In [Ali *et al.* 2008], QPBO was employed to solve the resulting pairwise model, nevertheless, only third-order potentials were tested in the experiments. A better reduction method that generally produces fewer non-submodular components was proposed in [Kolmogorov & Zabih 2004], in order to construct s-t graph for a third-order binary MRF. This reduction method was studied from an algebraic viewpoint in [Freedman & Drineas 2005] and led to some interesting conclusions towards extending this method to models of an arbitrary order. Based on these works, [Ishikawa 2009, Ishikawa 2011] proposed a generalized technique that can reduce any higher-order binary MRF into a pairwise one, which can then be solved by QBPO. The same concept was extended in [Ishikawa 2009, Ishikawa 2011] to deal with multi-label MRFs by using fusion moves [Lempitsky *et al.* 2010]. Very recently, aiming to obtain a pairwise model that are as easy as possible to solve, [Gallagher *et al.* 2011] proposed to approach order reduction as a optimization problem, where different factors are allowed to choose different reduction methods towards optimizing an objective function defined using a special graph (referred to as *order reduction inference graph*).

Graph-cuts methods have also been considered to cope either with specific visual perception problems or certain classes of higher-order models. For example, [Kohli *et al.* 2007, Kohli *et al.* 2009b] characterized a class of higher-order potentials (*i.e.*, $\mathcal{P}^n$ Potts model) for which the optimal expansion and swap moves can be computed efficiently in polynomial time, and proposed an efficient graph-cuts-based method for solving such models. Such a technique was further extended in [Kohli *et al.* 2008a, Kohli *et al.* 2009a] to a wider class of higher-order models (*i.e.*, robust $\mathcal{P}^n$ model). Graph-cuts-based approaches were also proposed [Ladicky *et al.* 2010a, Ladicky *et al.* 2011] and in [Ladicky *et al.* 2010a, Ladicky *et al.* 2011] to perform inference in their higher-order MRFs with global potentials that encode "co-occurrence statistics" and/or "label costs". Despite the fact that such

[24] Here, we consider binary higher-order MRFs and their energy functions can be represented in form of *pseudo-Boolean functions* [Boros & Hammer 2002].

methods were designed for a limited range of problems, they better capture the characteristics of the problems and are able to solve the problems relatively efficiently (*e.g.*, they often cannot be solved by a general inference methods).

Belief-propagation-based Methods

As we mentioned in section 2.2.2, the factor graph representation of MRFs enables the extension of classic min-sum belief propagation algorithm to higher-order cases. Hence loopy belief propagation in factor graphs provides a straightforward way to deal with inference in higher-order MRFs. Such an approach was employed in [Lan *et al.* 2006] to solve their higher-order Fields-of-Experts model.

A practical problem for propagating messages in higher-order MRFs is that the complexity increases exponentially with respect to the highest order among all cliques. Various techniques have been proposed to accelerate the belief propagation in special families of higher-order potentials. For example, the use of *distance transform* techniques [Borgefors 1986, Felzenszwalb & Huttenlocher 2006] significantly improves the efficiency of the message passing process in [Lan *et al.* 2006]. [Potetz 2007, Potetz & Lee 2008] and [Tarlow *et al.* 2010] proposed efficient message passing algorithms for some families of potentials such as *linear constraint potentials* and *cardinality-based potentials*. Recently, the max-product message passing was accelerated in [Mcauley & Caetano 2011] by exploiting the fact that a clique potential often consists of a sum of potentials each of which involves only a sub-clique of variables. The expected time of the message passing was further reduced in [Felzenszwalb & Mcauley 2011].

Dual Methods

The LP relaxation formulation in Eq. 2.25 can be generalized to the cases of higher-order MRFs. Such a generalization was studied in [Werner 2008, Werner 2010], where *min-sum diffusion* [Kovalevsky & Koval 1975] was adopted to achieve a method for optimizing the energy of higher-order MRFs, which is referred to as *n-ary min-sum diffusion*[25].

The *Dual-decomposition* framework [Bertsekas 1999, Komodakis *et al.* 2007b], which has been presented in section 2.2.3, can also be adopted to deal with higher-order MRFs. This was demonstrated in [Komodakis & Paragios 2009], where inference algorithms were introduced for solving a wide class of higher-order potential referred to as *pattern-based potentials*[26].

[25] The method was originally called *n-ary max-sum diffusion* in [Werner 2008, Werner 2010] due to the fact that a maximization of objective function was considered.

[26] For example, $\mathcal{P}^n$ Potts model [Kohli *et al.* 2009b] is a sub-class of *pattern-based potentials*.

Lastly, we note that the exploitation of the sparsity of potentials is explicitly or implicitly employed in many of the above higher-order inference methods. In this direction, [Rother *et al.* 2009] proposed a compact representation for "sparse" higher-order potentials (except a very small subset, the labelings are almost impossible so as to have the same high energy) to convert a higher-order model into a pairwise one so that pairwise MRF inference methods such as graph cuts can be employed to solve the problem. Due to the "sparseness", only a small number of auxiliary variables are required for the order reduction process. In the same line of research, [Kohli & Pawan Kumar 2010] studied and characterized some families of higher-order potentials (*e.g.*, $\mathcal{P}^n$ Potts model [Kohli *et al.* 2009b]) that can be represented compactly as upper or lower envelopes of linear functions. Furthermore, it was demonstrated that these higher-order models can be converted into pairwise models with the addition of a small number of auxiliary variables.

2.3 Conclusion

In order to conclude this chapter, let us first recall to the reader that graphical models, in particular Markov Random Fields and discrete optimization have been a dominant research direction in computer vision for the past decade. The main stream referred to pairwise formulations, where their use was mostly motivated from computational efficiency. In the recent years, we have witnessed significant progress with regards to the optimization of MRFs, in particular higher-order MRFs, which was the main driving force of this thesis. The use of distributed graphical models and the master-slave decomposition will be the concepts being studied and developed towards addressing some of the most fundamental problems of low, mid and high-level vision.

Chapter 3

Segmentation, Depth Ordering and Multi-object tracking

In this chapter, we aim to jointly and simultaneously solve segmentation, multi-object tracking and depth ordering from monocular video sequences using a unified graph-based framework. To this end, we first propose a joint 2.5D layered model where top-down object-level and bottom-up pixel-level representations are seamlessly combined through local constraints which involve only pairs of variables. Then based on such a layered model, we propose a graphical-model formulation, where all the observed and hidden variables of interest such as image intensities, states of pixels (index of the associated object and relative depth) and of objects (motion parameters[1] and relative depth) are jointly modeled within a single pairwise MRF. Finally, through minimizing the MRF energy, we simultaneously segment, track and sort by depth the objects. Promising experimental results demonstrate the potential of this framework and its robustness to image noise, cluttered background, moving camera and background, and even complete occlusions.

3.1 Introduction

Image segmentation and object tracking are among the most fundamental and active research topics in the computer vision community. They often serve as low and mid-level cues in numerous applications such as video surveillance, action recognition, robot navigation, medical imaging and human-machine interaction.

[1] Here, *motion parameters* are referred to as all the parameters controlling the shape of an object, such as global pose parameters (location, scale, rotation) and other parameters for characterizing the shape variation.

Image segmentation aims at grouping pixels into meaningful components/regions or delineating boundaries between different regions. Hence, segmentation methods can be classified into two categories: edge-based and region-based. In the first category, one seeks the region boundaries that do often correspond to visual discontinuities. *Active contours* (including *snakes* [Kass *et al.* 1988], *geodesic active contours* [Caselles *et al.* 1997] and their implicit *level set* variants techniques [Osher & Fedkiw 2002, Osher & Paragios 2003]) are popular methods. The central idea is to evolve and propagate an initial curve towards the desired region boundaries under the influence of image forces while being constrained from internal ones. In the second category, pixels are grouped together according to their visual properties and spatial relationships, either through clustering, continuous or discrete methods. Mean-shift [Comaniciu & Meer 2002] is a typical example of clustering method that aims at grouping together pixels with the same chromaticity characteristics. Examples of the continuous methods include *Mumford-Shah* [Mumford & Shah 1989], *Chan-Vese* functionals [Chan & Vese 2001] and *geodesic active regions* [Paragios & Deriche 2002], usually solved via level set approaches [Osher & Fedkiw 2002, Osher & Paragios 2003]. Regarding discrete methods, graph-based methods have been quite popular, like *normalized cut (NCut)* [Shi & Malik 2000], *isoperimetric cut (IsoCut)* [Grady & Schwartz 2006] and more recent MRF-based segmentation techniques are the current state of the art (*e.g.*, [Rother *et al.* 2004, Boykov & Funka-Lea 2006]). In such a context, each pixel is endowed a label (discrete variable) denoting the segment it belongs to and the pixel labeling can be efficiently achieved via discrete optimization methods. Such an approach inherits the advantage of being less susceptible - over continuous methods and active contours - to local minima. As a very important sub-problem, segmenting a specific object category such as human body and organs (*e.g.*, [Cootes *et al.* 1995, Kohli *et al.* 2008b, Besbes *et al.* 2009]), have also raised lots of attentions. Among existing methods, prior information on the shape of the specific class is usually combined within the approach towards improving significantly the quality of segmentation (*e.g.*, [Freedman & Zhang 2005, Huang *et al.* 2004, Kohli *et al.* 2008b]).

Object tracking aims at locating moving objects in consecutive frames of a video sequence. The representation of an object of interest usually consists of a shape (that can be fairly simple or fairly complex) and an appearance models. Shape representations encompass basic geometric shapes (*e.g.*, rectangle, ellipse [Comaniciu *et al.* 2000, Comaniciu *et al.* 2003]), complex parametric shape representations [Balan & Black 2006, Kohli *et al.* 2008b, de La Gorce *et al.* 2008] and part-based models [Sudderth *et al.* 2004a, Felzenszwalb & Huttenlocher 2005, Sigal & Black 2006a], *etc*. Such geometric priors are often combined with image-based similarities of the object appearance in time to estimate the configuration of the object in each frame. Furthermore, dynamical system can also be adopted to encode information on the object trajectory properties. *Kalman filters*

[Kalman 1960, Gelb 1974], the *mean-shift* algorithm [Comaniciu *et al.* 2000] and the *condensation* [Isard & Blake 1998] are the most popular methods for single object tracking. The transition from single object tracking to the multi-object brings in the inevitable task of efficiently dealing with the interactions between objects, in particular occlusions during the movement.

3.1.1 Joint Segmentation and Tracking

Image segmentation and object tracking are two complementary tasks. More and more researchers aim to jointly and simultaneously solve them in order to improve their performances. In the MRF segmentation literature, one can cite for example the works of [Huang *et al.* 2004, Freedman & Zhang 2005, Pawan Kumar *et al.* 2005], where object shape priors were considered in the context of MRF segmentation. These approaches determine the segmentation and the shape estimation using *coordinate descent* or EM-style optimizations where the global objective function is minimized with respect to the shape model parameters and pixel labeling, alternately.

This concept was further enhanced in [Kohli *et al.* 2008b], where articulated object tracking and MRF segmentation were coupled within an objective function. Such a formulation involves both pixel labeling and the pose parameters. In order to solve this complex inference problem due to the fact that discrete and continuous variables are present, they proposed a combination of continuous and discrete optimizations: for a given pose configuration, dynamic binary graph cuts [Kohli & Torr 2005] are used to determine the minimum of the function with respect to the pixel labeling. Once a binary segmentation map is determined, a gradient-free local search (via *Powell* minimization algorithm [Press *et al.* 1988]) is performed to determine the pose parameters. Such a promising approach in the context of single object tracking is not suited for multi-object tracking unless proper handling of occlusions between objects is introduced. The direct extension of this approach to the case of multiple objects would cause *evidence over-counting* problem (*i.e.*, associating a pixel to more than one object). [Malcolm *et al.* 2007] proposed to solve multi-object tracking and image segmentation via multi-label graph cuts [Boykov *et al.* 2001]. More specifically, template shapes of the objects, with the position parameters predicted from previous frames, are used as shape priors to perform multi-label MRF image segmentation with graph cuts. Then the object positions are re-estimated using the segmented regions. The use of multi-label segmentation helps in avoiding evidence over-counting but is insufficient to ensure robustness with respect to (moderate and severe) occlusions that would require occlusion reasoning.

However, it is important to note that such combined methods perform in general better than the ones solving the problems sequentially (*e.g.*, [Huang & Essa 2005, Yang *et al.* 2005, Agarwal & Triggs 2004]). Such an observation is evident when more challenging conditions such as image noise and cluttered background are present.

3.1.2 Depth Ordering and Occlusion Handling

Multi-object tracking in monocular video sequences becomes a challenging computer vision problem, mostly due to the occlusions caused by the overlapping of objects along the line of view. In the recent years, substantial effort has been dedicated to deal with occlusions in the literature (*e.g.*, [Jepson *et al.* 2002, Huang & Essa 2005, Yang *et al.* 2005, Senior *et al.* 2006]). One essential issue of occlusion handling is how to process the data association so as to correctly explain the evidence from observed images. Another concern relates on how to account for the occluded parts of objects towards properly estimating the spatial configuration of objects in the cases where the objects are partially or completely occluded. However, it is not straightforward to take into account such objectives in a graphical-model formulation without introducing high-order cliques. [Sigal & Black 2006a] and [Sudderth *et al.* 2004a] proposed to combine binary visibility variables within graph-based tracking to perform occlusion reasoning. Occlusions are considered towards avoiding over-counting image support. Nevertheless, the formulations did not intrinsically guarantee that at least one object or the background has to be associated to a given pixel.

Depth notion and *layered models* were other alternatives that were widely used in the literature (*e.g.*, [Nitzberg & Mumford 1990, Wang & Adelson 1994, Darrell & Fleet 1995, Tao *et al.* 2000, Jojic & Frey 2001, Jepson *et al.* 2002, Winn & Blake 2004, Smith *et al.* 2004, Jackson *et al.* 2008, Pawan Kumar *et al.* 2008, Auvray *et al.* 2009, Sun *et al.* 2010]). Layered representations provide a compact spatial modeling by considering succinctly a region/object as a layer. A 2.5D representation, where *relative depth* is introduced to each layer, allows the use of depth ordering to perform visibility reasoning and occlusion handling in an explicit and rigorous way. One can cite for example the work [Jepson *et al.* 2002] which proposed to combine depth ordering to perform object tracking. However, an inevitable issue is how to efficiently perform the inference. In the previous (generative) approaches, the layers are usually strictly and totally ordered according to their relative depths. The use of such a depth ordering and other scene parameters lead to a high-order objective function which involves all the objects and cannot be factorized. Two kinds of methods have been used to optimize such a function. One is coordinate-descent or Expectation-maximization (EM) method such as the one proposed in [Jepson *et al.* 2002]. An alternative strategy considers depth ordering as a hyper-parameter of the whole formulation (*e.g.*, [Smith *et al.* 2004]). Then one can evaluate the optimum of the objective function for each possible ordering configuration by optimizing the function with respect to other parameters. The optimal solution of the combined problem corresponds then to the best one among all ordering candidates. Such an approach increases dramatically the complexity with respect to the number of objects, since the number of ordering configurations is the factorial of the number of objects.

3.1.3 Our Approach

In the remaining of this chapter, we aim to introduce a method that addresses in a sound and valid manner the multi-object tracking and segmentation problems while being efficient from computational viewpoint. In order to satisfy such requirements, let us define a number of desired principles for a combined multi-object segmentation and tracking approach as follows:

1. Proper integration of depth ordering within the whole tracking/segmentation formulation to modeling rightly and rigorously visibility and occlusion;

2. Joint and simultaneous estimation of all variables of interest (depth, motion parameters and pixel segmentation labels);

3. Integration within a single MRF towards taking advantage of generic MRF inference techniques (see chapter 2), which are less prone to be trapped in local minima than local search or EM-style techniques.

In order to meet the above conditions, the main theoretical challenge lies on the decomposition of the depth ordering into low-order interactions between variables, which then can be easily integrated with standard MRF-based segmentation and tracking components. Our first main contribution lies on a novel joint 2.5D layered image modeling, where only pairwise interactions can encode all necessary visibility constraints. Then based on such a modeling, we have achieved a unified MRF formulation to address the challenge of combining the segmentation and multi-object tracking with a rigorous visibility modeling (*i.e.*, depth ordering). The latent states of pixels (index of the associated object and relative depth) and of objects (motion parameters, relative depth) are integrated along with a principled way in the MRF. By minimizing the MRF energy, we simultaneously segment the image, estimate the motion parameters of the objects and sort by depth the objects.

3.1.4 Outline of the Chapter

The remainder of this chapter is organized as follows. We present in section 3.2 our joint 2.5D layered modeling, which is then transported into the MRF formulation for the integrated multi-object tracking, ordering and image segmentation in section 3.3. Experimental validation and some discussion compose section 3.4. Finally, we conclude the chapter in section 3.5.

3.2 Joint 2.5D Layered Modeling

The proposed joint 2.5D layered model consists of top-down object-level and bottom-up pixel-level representations which are combined in a principled way. Let us first introduce some basic assumptions with regard to the above two-level representations. We assume that the objects of interest have two following properties:

1. There is no *mutual occlusion*[2] (*e.g.*, object 1 partially occludes object 2 and is partially occluded by object 2). Thus, each object can be considered to be "flat" and modeled as a 2D shape, especially for the purpose of visibility modeling. We regard each object as well as the background as a layer[3].

2. The objects are opaque. Thus, one and only one object is visible at any location in the image plane.

3.2.1 Object-level Representation

Let us assume that we know that there are K objects of interest in a sequence of images. Thus, we use $\mathcal{V}_o = \{1, 2, \ldots, K\}$ to denote the index set of the objects (*i.e.*, the foreground layers). Furthermore, we model the background as a special object (*i.e.*, the background layer) and assign it an index "0". Finally, we define the extended object set $\mathcal{V}_s = \mathcal{V}_o \cup \{0\}$ which contains the indices of all the layers of the scene.

The spatial configuration of each object k $(k \in \mathcal{V}_o)$ consists of two components:

1. a 2D parametric shape model $\mathcal{M}_k(\theta_k)$ with motion parameters θ_k (*e.g.*, location, scale, rotation for the case of similarity transform), which characterizes the "horizontal" extent in the image plane;

2. a *relative depth*[4] d_k, which characterizes the "vertical" position in the layered hierarchy and is used to determine the occlusion relation between the objects. An object i can occlude another object j only if $d_i < d_j$. A detail discussion on the depth will be provided in section 3.2.1, in particular its domain of definition.

[2] A mathematical definition will be given in section 3.2.1.

[3] Hereafter, we use the term *layer* to refer to *object or background* for the purpose of conciseness.

[4] Towards simplifying the presentation of the framework, the term *relative depth* will be replaced by *depth* hereafter.

Regarding the background layer, we adopt two simple conditions in terms of shape and occlusion: (i) its 2D shape $\mathcal{M}_0(\theta_0)$ is always equal to the image domain[5]; and (ii) it lies behind (is occluded by) all the objects to be tracked[6], and thus its depth d_0 is greater than that of any object.

To conclude, the composite spatial parameter $\Gamma_k = (\theta_k, d_k)$ characterizes the spatial configuration of layer k ($k \in \mathcal{V}_s$), and the *object-level representation* can be denoted as a vector of spatial parameters:

$$\mathbf{\Gamma} = (\Gamma_k)_{k \in \mathcal{V}_s} = (\theta_k, d_k)_{k \in \mathcal{V}_s} \tag{3.1}$$

For sake of convenience for some presentation, we also reformulate $\mathbf{\Gamma}$ as $\mathbf{\Gamma} = (\boldsymbol{\theta}, \mathbf{d})$ where $\boldsymbol{\theta} = (\theta_k)_{k \in \mathcal{V}_s}$ and $\mathbf{d} = (d_k)_{k \in \mathcal{V}_s}$ denote the shape parameters and the depths of all the objects, respectively.

Relative Depth

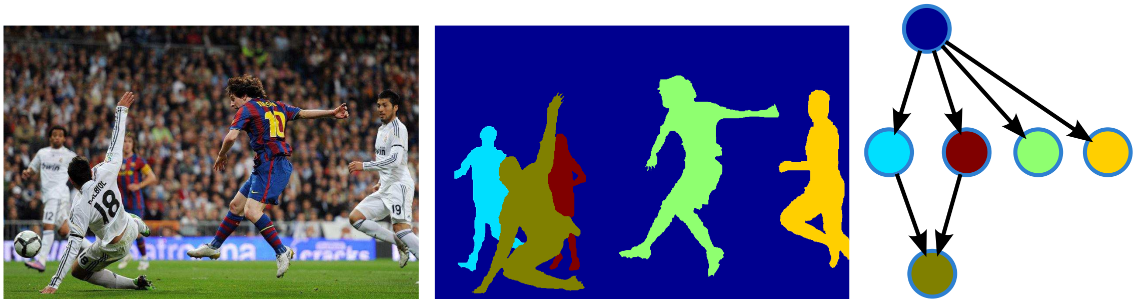
In previous generative frameworks (*e.g.*, [Jepson *et al.* 2002, Smith *et al.* 2004]), the depth configuration $\hat{\mathbf{d}} = (d_k)_{k \in \mathcal{V}_o}$ of the objects is usually considered as a permutation of $(0, 1, \ldots, |\mathcal{V}_o| - 1)$ and the depth of the background is a constant, *i.e.*, $d_0 = |\mathcal{V}_o|$. In such a representation, the objects and background are strictly and totally ordered by their depths (using the usual "less than" operator "<"), and thus the number of possible depth orderings is $|\mathcal{V}_o|!$. Such a depth modeling provides a sound theoretical approach to reasoning the visibility/occlusion. However, it is not compact and is somewhat "wasteful" in practice. This is due to the fact that the depth order between two objects is meaningful only when there is occlusion between them (including transitive occlusion via third objects) whereas the number of overlapping objects is usually much smaller than the total number of objects. Taking two completely visible objects for example, the relation between their depths can be arbitrary.

Here, we elaborate the modeling of the depth using the notion of *occlusion graph* $\mathcal{G}_o$ [Darrell & Fleet 1995] (Fig. 3.1), where a node i represents a layer i and a directed edge (i, j) indicates the relation between the two layers: layer i is occluded by layer j, which implies that layers i and j overlap. We start the presentation by defining *mutual occlusion* based on the occlusion graph as follows:

Definition 1. Mutual occlusion *is the occlusion relation between two or more layers that yields a cycle* $\mathbf{c} = (k_1, k_2, \ldots, k_1)$ *in the corresponding occlusion graph* $\mathcal{G}_o$.

[5] θ_0 is an abuse of notation (this variable is not needed), which is used for sake of clarity and consistency for the following presentation.

[6] The floating background, *i.e.*, those objects which are not tracked but may occlude the objects to be tracked will be discussed in section 3.4.2. However, it is not a limitation with regard to the proposed layered model.



(a) Original Image (b) Multi-label Segmentation (b) Occlusion Graph

Figure 3.1: Example of Occlusion Graph

Thus, the assumption that no mutual occlusion is present in the scene boils down to that the corresponding occlusion graph has no cycle, producing a *Directed Acyclic Graph* (DAG). Under this assumption, the structure of the occlusion graph can be fully characterized using the *parent-child* relation between nodes.

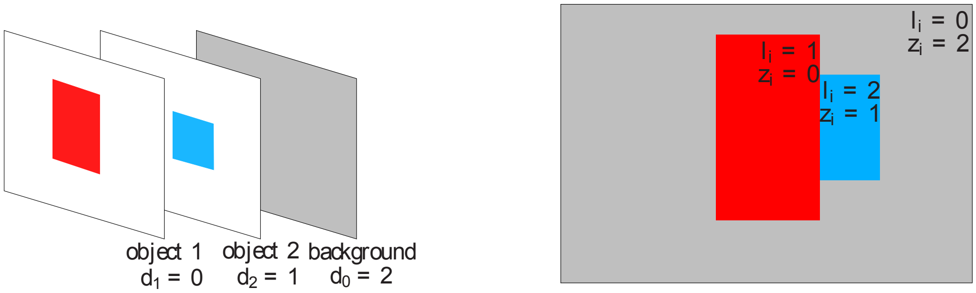
In fact, what we need to model towards visibility/occlusion reasoning is the structure of the occlusion graph, *i.e.*, the parent-child relation between nodes in cases without mutual occlusion. As the parent-child relation is a *partial order*, if we want to achieve the visibility reasoning purpose by associating a node i with a depth d_i, the only condition we need to guarantee is $d_i > d_j$ for any edge (i, j) in $\mathcal{G}_o$. Under this condition, we can correctly reason which layer is visible at a certain location among the overlapping layers using their depths. Hence, assuming that at most D ($D \leq K$) objects (not including the background) may overlap in an observed image[7], $D + 1$ depths are sufficient to model the visibility between the objects and the background. We define $\mathcal{D} = \{0, 1, 2, \ldots, D - 1\}$ as the set of all the possible depths for the objects, and "D" as the depth of the background, *i.e.*, $d_k \in \mathcal{D}$ $(k \in \mathcal{V}_o)$ and $d_0 = D$. Note that the number D of overlapping objects is usually rather small in real scene, where this modeling of depths leads to a much more compact space of the depth vector $\mathbf{d}$.

3.2.2 Pixel-level Representation

We assume that an observed image consists of N pixels, and use $\mathcal{V}_p = \{K + 1, K + 2, \ldots, K + N\}$ to denote the index set of the pixels[8]. Under the assumption on the opaqueness of objects, each pixel is to be assigned to one and only one layer. In order to model

[7] More formally, D is the length of the longest directed path in the occlusion graph $\mathcal{G}_o$.

[8] The pixels are indexed from $K + 1$ in order to avoid overlapping with the indices of objects and to be coherent with the indices of the corresponding nodes in the MRF formulation (see section 3.3).



(a) Object-level Representation

(b) Pixel-level Representation

Figure 3.2: Sketch Map of the Joint 2.5D Modeling

this, for each pixel i ($i \in \mathcal{V}_p$), we introduce a latent variable (named *pixel label*) l_i ($l_i \in \mathcal{V}_s$) which provides the index of the layer to which the pixel i is associated (*i.e.*, layer l_i is "visible" at pixel i). And thus $\mathbf{l} = (l_i)_{i \in \mathcal{V}_p}$ denotes the index of the associated layer for all the pixels, *i.e.*, the segmentation of the image.

In order to combine the object-level and pixel-level representations using only local pairwise constraints (section 3.2.3), we assign a *depth* z_i ($z_i \in \mathcal{D} \cup \{D\}$) to each pixel i. It represents the depth of the layer to which the pixel associates, *i.e.*, $z_i = d_{l_i}$.

To conclude, for each pixel i, the composite parameter $\Lambda = (l_i, z_i)$ characterizes the index and the depth of the associated layer. The *pixel-level representation* can be denoted as:

$$\mathbf{\Lambda} = (\Lambda)_{i \in \mathcal{V}_p} = (l_i, z_i)_{i \in \mathcal{V}_p} \tag{3.2}$$

3.2.3 Combination of the Two-level Representations

The multi-object segmentation, tracking and depth ordering problem boils down to the inference of the latent values of shape parameters $\boldsymbol{\theta}$, depths $\mathbf{d}$ and the pixel labels $\mathbf{l}$ in the above layered representations. In this section, we combine the two representations together and derive the conditions for a valid configuration $(\mathbf{\Gamma}, \mathbf{\Lambda})$ of the joint 2.5D layered model.

Let us first introduce, from a generative viewpoint, three types of visibility constraints between the object-level configuration $\mathbf{\Gamma}$ and the pixel-level configuration $\mathbf{\Lambda}$:

1. *Pixel Label Consistency* encodes constraints on the data association (*i.e.*, which layer is "visible" at a pixel i) within the top-down generative process. It imposes that a given pixel i is assigned to the layer having the smallest depth among the layers whose shapes are likely to project to this pixel. It can be formulated in a rigorous

mathematical form as follows:

$$\forall\, i \in \mathcal{V}_p \,,\, l_i = \underset{\{k|i\in\mathcal{M}_k(\theta_k),k\in\mathcal{V}_s\}}{\arg\min} d_k \tag{3.3}$$

where we recall that $\mathcal{M}_k(\theta_k)$ denotes the 2D shape model of object k with parameter θ_k, and we regard $\mathcal{M}_k \subset R^2$ as the union of the interior region and the boundary of object k.

2. *Object Depth Consistency* encodes constraints on the scene configuration $\mathbf{\Gamma}$ in order to guarantee that one and only one layer is "visible" at any pixel i (*i.e.*, to guarantee that *Pixel Label Consistency* is well defined, $\arg\min_{\{k|i\in\mathcal{M}_k(\theta_k),k\in\mathcal{V}_s\}} d_k$ should be singleton). We can formulate this as follows:

$$\begin{aligned} &\forall\, i \in \mathcal{V}_p \,,\, \exists \tilde{k} \in \{k|i \in \mathcal{M}_k(\theta_k) \,,\, k \in \mathcal{V}_s\} \\ &\text{s.t. } \forall\, k' \in \{k|i \in \mathcal{M}_k(\theta_k) \,,\, k \in \mathcal{V}_s\}\backslash\{\tilde{k}\} \,,\, d_{\tilde{k}} < d_{k'} \end{aligned} \tag{3.4}$$

3. *Pixel Depth Consistency* encodes constraints between the depths in object-level and pixel-level towards assigning consistent labels. It imposes that the depth of a pixel i has to be equal to the depth of the layer which is "visible" at this pixel, and can be formulated as:

$$\forall\, i \in \mathcal{V}_p \,,\, z_i = \min_{\{k|i\in\mathcal{M}_k(\theta_k),k\in\mathcal{V}_s\}} d_k \tag{3.5}$$

The combination of *Pixel Label*, *Object Depth* and *Pixel Depth* consistencies (Eqs. 3.3 $\sim$ 3.5) guarantees a valid configuration $(\mathbf{\Gamma}, \mathbf{\Lambda})$ for the joint 2.5D layered model (Fig. 3.2). Such a condition can be formulated in a distributed manner so that the visibility reasoning and the validation of the model configuration are performed through local pairwise constraints between a layer and a pixel.

$$\forall\, i \in \mathcal{V}_p, \quad \mathcal{A}_1 \wedge \mathcal{A}_2 \wedge \mathcal{A}_3 \Leftrightarrow \bigwedge_{k\in\mathcal{V}_s} (\mathcal{C}_{1k} \wedge \mathcal{C}_{2k} \wedge \mathcal{C}_{3k}) \tag{3.6}$$

with:

$$\begin{cases} \mathcal{A}_1: & l_i = \arg\min_{\{k|i\in\mathcal{M}_k(\theta_k),k\in\mathcal{V}_s\}} d_k \\ \mathcal{A}_2: & z_i = \min_{\{k|i\in\mathcal{M}_k(\theta_k),k\in\mathcal{V}_s\}} d_k \\ \mathcal{A}_3: & \exists \tilde{k} \in \{k|i \in \mathcal{M}_k(\theta_k) \,,\, k \in \mathcal{V}_s\} \text{ s.t.} \\ & \forall\, k' \in \{k|i \in \mathcal{M}_k(\theta_k) \,,\, k \in \mathcal{V}_s\}\backslash\{\tilde{k}\} \,,\, d_{\tilde{k}} < d_{k'} \\ \mathcal{C}_{1k}: & \neg((l_i = k) \wedge (z_i \neq d_k)) \\ \mathcal{C}_{2k}: & \neg((l_i = k) \wedge (i \notin \mathcal{M}_k(\theta_k))) \\ \mathcal{C}_{3k}: & \neg((l_i \neq k) \wedge (z_i \geq d_k) \wedge (i \in \mathcal{M}_k(\theta_k))) \end{cases} \tag{3.7}$$

Proof. Let $\mathcal{A} = \mathcal{A}_1 \wedge \mathcal{A}_2 \wedge \mathcal{A}_3, \mathcal{C} = \bigwedge_{k \in \mathcal{V}_s}(\mathcal{C}_{1k} \wedge \mathcal{C}_{2k} \wedge \mathcal{C}_{3k})$ and $\mathcal{O}_i = \{k | i \in \mathcal{M}_k(\theta_k), k \in \mathcal{V}_s\}$.

"⇒": We first prove that for a pixel i ($\forall\, i \in \mathcal{V}_p$), "$\mathcal{A}$ is true" then "$\mathcal{C}$ is true" using *Reduction to the absurd*:

1. Assuming that $\exists \tilde{k} \in \mathcal{V}_s$ s.t. $\mathcal{C}_{1\tilde{k}}$ is false, then $l_i = \tilde{k}$ and $z_i \neq d_{\tilde{k}}$. But according to $\mathcal{A}_1$, $\mathcal{A}_2$, $z_i = d_{l_i} = d_{\tilde{k}}$. So the assumption is wrong, *i.e.*, $\forall\, k \in \mathcal{V}_s, \mathcal{C}_{1k}$ is true.

2. Assuming that $\exists \tilde{k} \in \mathcal{V}_s$ s.t. $\mathcal{C}_{2\tilde{k}}$ is false, then $l_i = \tilde{k}$ and $i \notin \mathcal{M}_{\tilde{k}}(\theta_{\tilde{k}})$. But according to $\mathcal{A}_1$, $l_i \in \mathcal{O}_i$ then $\tilde{k} \in \mathcal{O}_i$, *i.e.*, $i \in \mathcal{M}_{\tilde{k}}(\theta_{\tilde{k}})$. So the assumption is wrong, *i.e.*, $\forall\, k \in \mathcal{V}_s, \mathcal{C}_{2k}$ is true.

3. Assuming that $\exists \tilde{k} \in \mathcal{V}_s$ s.t. $\mathcal{C}_{3\tilde{k}}$ is false, then $l_i \neq \tilde{k}$, $z_i \geq d_{\tilde{k}}$ and $i \in \mathcal{M}_{\tilde{k}}(\theta_{\tilde{k}})$. So $\tilde{k} \in \mathcal{O}_i$. And according to $\mathcal{A}_1$ and $\mathcal{A}_2$, $d_{l_i} = z_i \geq d_{\tilde{k}}$. But according to $\mathcal{A}_1$ and $\mathcal{A}_3$, $d_{l_i} < d_{k'}$ ($\forall\, k' \in \mathcal{O}_i \backslash \{l_i\}$). So the assumption is wrong, *i.e.*, $\forall\, k \in \mathcal{V}_s, \mathcal{C}_{3k}$ is true.

"⇐": Now we prove that for a pixel i ($\forall\, i \in \mathcal{V}_p$), "$\mathcal{C}$ is true" then "$\mathcal{A}$ is true":

$$\begin{aligned} \mathcal{C} =& (\bigwedge_{k \in \mathcal{V}_s} \mathcal{C}_{1k}) \wedge (\bigwedge_{k \in \mathcal{V}_s} \mathcal{C}_{2k}) \wedge (\bigwedge_{k \in \mathcal{V}_s} \mathcal{C}_{3k}) \\ =& \underbrace{(\neg \bigvee_{k \in \mathcal{V}_s} (\neg \mathcal{C}_{1k}))}_{\mathcal{C}_1} \wedge \underbrace{(\neg \bigvee_{k \in \mathcal{V}_s} (\neg \mathcal{C}_{2k}))}_{\mathcal{C}_2} \wedge \underbrace{(\neg \bigvee_{k \in \mathcal{V}_s} (\neg \mathcal{C}_{3k}))}_{\mathcal{C}_3} \end{aligned} \tag{3.8}$$

$$\begin{aligned} \mathcal{C}_1 &\Leftrightarrow \nexists k \in \mathcal{V}_s, (l_i = k) \wedge (z_i \neq d_k) \\ &\Rightarrow d_{l_i} = z_i \qquad (3.9) \\ \mathcal{C}_2 &\Leftrightarrow \nexists k \in \mathcal{V}_s, (l_i = k) \wedge (i \notin \mathcal{M}_k(\theta_k)) \\ &\Rightarrow l_i \in \mathcal{O}_i \qquad (3.10) \\ \mathcal{C}_3 &\Leftrightarrow \nexists k \in \mathcal{V}_s, (l_i \neq k) \wedge (z_i \geq d_k) \wedge (i \in \mathcal{M}_k(\theta_k)) \\ &\Rightarrow \forall\, k' \in \mathcal{O}_i \backslash \{l_i\}, z_i < d_{k'} \qquad (3.11) \end{aligned}$$

1. (3.9) and (3.11) ⇒ $\forall\, k' \in \mathcal{O}_i \backslash \{l_i\}, d_{l_i} < d_{k'}$. And according to (3.10), $l_i \in \mathcal{O}_i$. So $\mathcal{A}_1$ and $\mathcal{A}_3$ are true.

2. (3.9) and $\mathcal{A}_1$ (has been proved to be true) ⇒ $z_i = d_{l_i} = d_{\arg\min_{k \in \mathcal{O}_i} d_k} = \min_{k \in \mathcal{O}_i} d_k$, *i.e.*, $\mathcal{A}_2$ is true.

□

Interpretation of the local constraints

Let us now proceed with contextual interpretation of the derived constraints.

1. Keeping $\mathcal{C}_{1k}$ true imposes that: the depth of pixel i should be equal to the depth of layer k if it associates to the layer k.

2. Keeping $\mathcal{C}_{2k}$ true imposes that: a pixel i can associate to layer k only when it is occupied by the shape of layer k.

3. Keeping $\mathcal{C}_{3k}$ true imposes that: if a pixel i is occupied by the shape of layer k, it can associate to a layer other than k only when the depth of pixel i is strictly smaller than the depth of layer k.

Given the equivalence presented in Eq. 3.6, the satisfaction of the local conditions on the right-side for each pixel ensures that a pixel i will be explained once and only once by the object which is supposed to be visible at pixel i. One can now integrate these constraints with support coming from the image towards segmentation, depth ordering and multi-object tracking. Such an integration can be performed using a pairwise MRF model. This is doable because the model satisfaction conditions can be mapped to pairwise interactions, while image support can be encoded through singleton potentials.

3.3 Markov Random Field Formulation

The proposed MRF model consists of two types of nodes (Fig. 3.3). The first are *object* nodes corresponding to the objects to be tracked, and the second are *pixel* nodes corresponding to the image pixels. The index set of the nodes is denoted by $\mathcal{V} = \mathcal{V}_o \cup \mathcal{V}_p$, where $\mathcal{V}_o$ and $\mathcal{V}_p$ correspond to the two types of nodes, respectively[9]. Each node i ($i \in \mathcal{V}$) is associated with a latent random variable X_i which denotes the configuration of the corresponding object/pixel and takes a value x_i from its candidate set $\mathcal{X}_i$.

- **Pixel node:** The latent random variable X_i ($i \in \mathcal{V}_p$) is composed of the index of the associated layer and the depth, *i.e.*, $x_i = (l_i, z_i)$. We define the configuration space of pixel node i as: $\mathcal{X}_i = (\mathcal{V}_o \times \mathcal{D}) \cup \{(0, D)\}$. Note that if a pixel is labeled as "background" (*i.e.*, $l_i = 0$), its depth is deterministic (*i.e.*, $z_i = D$).

- **Object node:** The latent random variable X_k ($k \in \mathcal{V}_o$) consists of the motion parameters and the depth, *i.e.*, $x_k = (\theta_k, d_k)$. We use $\mathcal{X}_k = \Theta_k \times \mathcal{D}$ to denote the configuration space of object node k, where Θ_k denotes the motion parameter space.

[9] Due to the one-to-one mapping between the object node and the object, the pixel node and the pixel, in this section, we don't distinguish pixel node and pixel, object node and object.

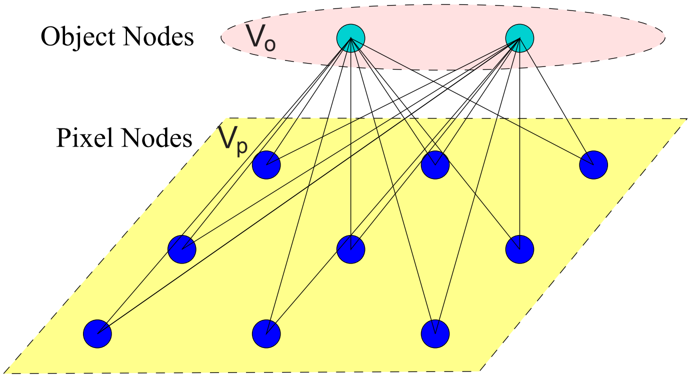


Figure 3.3: MRF Model (Example for Two Tracked Objects)

The whole MRF comprises a latent random variable vector $\mathbf{X} = (X_i)_{i \in \mathcal{V}}$. We use $\mathbf{x} = (x_i)_{i \in \mathcal{V}}$ to denote the configuration of the MRF and $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_{|\mathcal{V}|}$ its space, *i.e.*, $\mathbf{x} \in \mathcal{X}$.

In order to introduce the prior on the joint layered model presented in section 3.2, the object nodes are connected with all the pixel nodes (Fig. 3.3). These edges compose the edge set $\mathcal{E}$ of the MRF, *i.e.*, $\mathcal{E} = \{(k,i) | k \in \mathcal{V}_o, i \in \mathcal{V}_p\}$. We can also introduce interactions/dependencies on the labels of the pixel nodes (in particular with respect to the segmentation) through conventional 4-neighborhood or 8-neighborhood systems [Boykov & Funka-Lea 2006], which will be discussed in section 3.4.3.

The energy of the MRF with a configuration $\mathbf{x}$ is defined as a sum of singleton potentials and pairwise potentials:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \phi_i(x_i) + \sum_{\{k,i\} \in \mathcal{E}} \psi_{k,i}(x_k, x_i) \tag{3.12}$$

3.3.1 Singleton Potential

There are two types of singleton potentials, one referring to the pixel nodes and the other referring to the object nodes. They are used mainly to encode the intensity evidence coming from the observed image and object motion priors from one frame to the next.

Pixel Singleton Term

Like most of existing MRF segmentation approaches (*e.g.*, [Boykov & Funka-Lea 2006, Kohli *et al.* 2008b]), we use pixel singleton potential $\phi_i(x_i)$ $(i \in \mathcal{V}_p)$ to introduce the data likelihood, which imposes penalties for assigning l_i to pixel i and is defined as:

$$\phi_i(x_i) = -\log \Pr(\mathbf{I}_i | \mathbf{H}_{l_i}) \tag{3.13}$$

where $\mathbf{I}_i$ denotes the intensity/color (*e.g.*, RGB value) of pixel i, and $\mathbf{H}_k$ $(k \in \mathcal{V}_s)$ denotes the intensity/color distribution for layer k. We can model the color distribution for each layer using existing standard approaches such as a Gaussian mixture, a kernel-based approximation (*e.g.*, Parzen windows) of the distribution and outcome of linear or non linear classification techniques (*e.g.*, *Boosting* algorithms [Schapire 1990, Freund & Schapire 1997, Schapire 2001], *Randomized Forests* [Breiman 2001], *Support Vector Machines (SVMs)* [Boser *et al.* 1992, Cortes & Vapnik 1995, Muller *et al.* 2001]).

Object Singleton Term

The singleton potential for object node k encodes the prior preference on its spatial configuration x_k and can be defined as:

$$\phi_k^{(t)}(x_k^{(t)}) \quad = \quad \alpha_1 \cdot \rho(\theta_k^{(t)}, \hat{\theta}_k^{(t)}) + \alpha_2 \cdot d_k^{(t)} \tag{3.14}$$

where $\alpha_1 > 0$ and $\alpha_2 > 0$ are the weights for the corresponding terms, $\hat{\theta}_k^{(t)}$ is the predicted configuration of θ_k for instant t, and $\rho(\theta_k^{(t)}, \hat{\theta}_k^{(t)})$ denotes certain distance measure (*e.g.*, Euclidean distance) between $\theta_k^{(t)}$ and $\hat{\theta}_k^{(t)}$ which penalizes the deviation of the estimated configuration from the predicted one.

The first term imposes to certain extend temporal consistency for the motion. In particular, it can help to determine the motion of an object in cases where there is not enough visual information for the motion estimation (*e.g.*, the object is completely occluded by another object during a period). The choice of prediction model or dynamical system for $\hat{\theta}_k^{(t)}$ is independent from this framework and one can choose an off-the-shelf predictor.

The second term is used to eliminate arbitrary depth choices in case of depth ambiguities by favoring the smallest possible depth. In the absence of this term, different depths might yield the same MRF energy while all being valid solutions. One obvious example is the case of an object having no occlusion with any other object, as it can take any possible depth. However, removal of this term will not impact the performance of tracking and segmentation.

3.3.2 Pairwise Potential

The pairwise potential between an object and a pixel is used to model the prior on the layered model. For an edge (k, i) ($k \in \mathcal{V}_o$ and $i \in \mathcal{V}_p$), the pairwise potential $\psi_{k,i}(x_k, x_i)$ is defined as:

$$\psi_{k,i}(x_k, x_i) = \psi_{k,i}^{(1)}(x_k, x_i) + \psi_{k,i}^{(2)}(x_k, x_i) + \psi_{k,i}^{(3)}(x_k, x_i) \tag{3.15}$$

where $\psi^{(1)}_{k,i}$, $\psi^{(2)}_{k,i}$ and $\psi^{(3)}_{k,i}$ are the penalties for the cases where $\mathcal{C}_{1k} =$ false, $\mathcal{C}_{2k} =$ false and $\mathcal{C}_{3k} =$ false, respectively (see Eq. 3.7 for $\mathcal{C}_{1k}$, $\mathcal{C}_{2k}$ and $\mathcal{C}_{3k}$):

$$\begin{cases} \psi^{(1)}_{ki}(x_k, x_i) & = & \gamma_1 \cdot [\neg \mathcal{C}_{1k}] \\ \psi^{(2)}_{ki}(x_k, x_i) & = & \gamma_2 \cdot dist(i, \mathcal{M}_k(\theta_k)) \cdot [\neg \mathcal{C}_{2k}] \\ \psi^{(3)}_{ki}(x_k, x_i) & = & \gamma_3 \cdot dist(i, \mathcal{M}^c_k(\theta_k)) \cdot [\neg \mathcal{C}_{3k}] \end{cases} \tag{3.16}$$

where $\gamma_1 > 0$, $\gamma_2 > 0$ and $\gamma_3 > 0$ are the weights for the corresponding penalties, $\mathcal{M}^c_k(\theta_k)$ denotes the complement of $\mathcal{M}_k(\theta_k)$, Iverson Bracket $[\cdot]$ is defined as: for a statement S, $[S] = 1$ if S is true and 0 otherwise, and $dist(i, \mathcal{M})$ denotes the distance function [Osher & Fedkiw 2002] which is defined as the minimum Euclidean distance between the geometric shape corresponding to $\mathcal{M}$ and the spatial position $loc(i)$ of pixel i in the image:

$$dist(i, \mathcal{M}) = \min_{j \in \mathcal{M}} \|loc(i) - loc(j)\| \tag{3.17}$$

Since $dist(i, \mathcal{M}) = 0$ (if $i \in \mathcal{M}$), the pairwise potential defined in Eq. 3.15 and 3.16 can be reformulated more concisely as:

$$\begin{aligned} \psi_{k,i}(x_k, x_i) = {} & \gamma_1 \cdot [l_i = k] \cdot [z_i \neq d_k] \\ & + \gamma_2 \cdot [l_i = k] \cdot dist(i, \mathcal{M}_k(\theta_k)) \\ & + \gamma_3 \cdot [l_i \neq k] \cdot [z_i \geq d_k] \cdot dist(i, \mathcal{M}^c_k(\theta_k)) \end{aligned} \tag{3.18}$$

Instead of giving an infinite penalty to any case where a statement in Eq. 3.7 is false, we set $\psi^{(1)}_{ki}$ to be constant, $\psi^{(2)}_{ki}$ and $\psi^{(3)}_{ki}$ to be distance penalties. This is motivated by the fact that, in general, shape models are not exact: the closer a position is to the center of shape, the higher is the degree of certainty of being in the projection. Such a penalty yields an elastic force and can guide both object tracking and image segmentation.

Using the MRF model defined above, we can now simultaneously perform segmentation, depth ordering and multi-object tracking, which is formulated as the inference of those latent random variables through a minimization over the MRF energy:

$$\mathbf{x}^{\text{opt}} = \arg\min_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}) \tag{3.19}$$

3.4 Experimental Results

3.4.1 Experimental Setting

In order to validate the proposed framework, we have considered several challenging video sequences of hundreds of frames each, where noise, cluttered background, moving camera and background, and/or even complete occlusions are present.

A weak geometric prior was considered, which is a bounding box (except for *Shell Game* sequence, where the geometric prior is the manually delineated contour of each object in the first frame.). The motion parameters θ_k of each object correspond to the position, scale and rotation angle around the shape's center of mass. The position space is defined as the support (or: lattice) of the image. The rotation angle space is defined as the set $\mathbb{Z}$ of all integers. The scale factor space is defined as $\{s|s = 1.05^n, n \in \mathbb{Z}\}$. We process an observed video sequence frame by frame. Thus in practice, the search space for the current frame is in the vicinity of the previous motion parameter vector, due to the fact that the motion between two successive frames is expected to be small. For the distance measure function ρ in the object singleton potential (Eq. 3.14), we used Euclidean distance for simplification. Such a setting is combined with a linear predictor where the estimated motion parameter vector for the current frame is used to predict that of the next frame, *i.e.*, $\hat{\theta}_k^{(t)} = \theta_k^{\text{opt},(t-1)}$ $(k \in \mathcal{V}_o)$. We chose to use such a simple motion predictor in the experiments, in order to diminish the effect of the predictor in the occlusion handling and thus to sufficiently demonstrate the potential of our formulation.

For the visual appearance term, we distinguish the case of static background from that of dynamic background. In the first case, using the manual delineation of the objects in the first frame, a Gaussian mixture is considered towards modeling the color distribution of each object. The color of the background, either is globally modeled as a Gaussian mixture (*Box* and *Shell Game* sequences), or is modeled using a pixelwise model (*Pedestrian Sequence 1*), *i.e.*, each pixel's color is modeled using a Gaussian distribution whose mean and variance are learned from a sequence of background images [Stauffer & Grimson 1999]. The case of dynamic background (*Pedestrian Sequences 2 and 3*) is treated differently. Given the manual segmentation of the first frame, a non-parametric Parzen windows approximation is used to model the color distribution of each layer. The color model for the background is updated for the next frame using the segmentation result of the current frame, while those of the objects are kept constant.

There are two components still to be addressed, the motion parameter sampling and the parameter setting for the weights of the MRF's energy. We adopt a sparse sampling strategy [Glocker *et al.* 2008a], where $\theta_k^{(t)}$ is sampled uniformly along each main axis plus the two diagonal directions of the translation centered at the predicted value $\hat{\theta}_k^{(t)}$, plus $\hat{\theta}_k^{(t)}$ itself to compose the set Θ_k of motion parameter candidates. In order to mitigate inaccuracy of the solution due to the fact that the sampling is sparse, we iterate by resampling at each iteration around the solution found in the previous iteration. According to the roles of the energy terms, we set the parameters as follows: we adjust and fix γ_2 by trial and error on the first few frames. It is different from one sequence to another since the color statistics and/or the color model may be different. The rest are set as: $\gamma_1 = 50\gamma_2$ and $\alpha_1 = \alpha_2 = \gamma_3 = \gamma_2$.

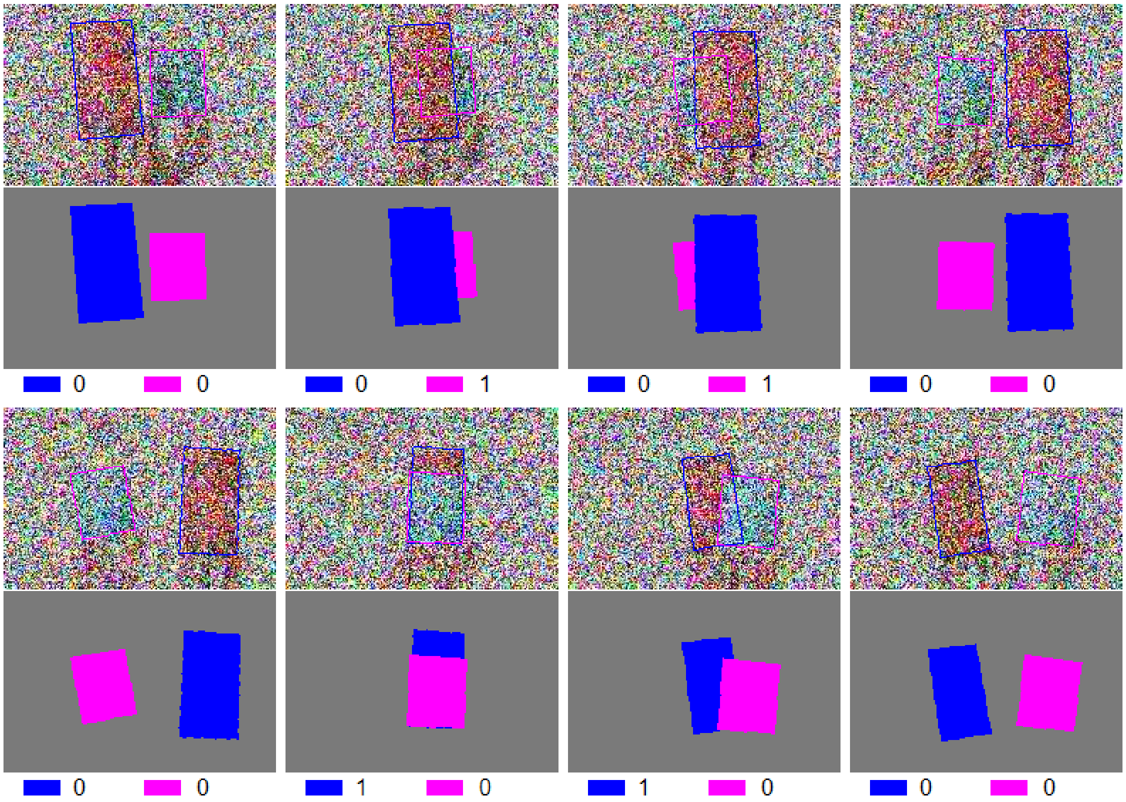


Figure 3.4: Experimental Results for *Box Sequence*. The first line of each sub-figure is the tracking result, where we draw the shape contours of the objects with the estimated motion parameters. The second line is the segmentation result. The third line presents the estimated depths of the objects. We use different colors to distinguish the objects. Same for the rest results in this chapter.

The MRF energy in Eq. 3.19 can be optimized using standard MRF-MAP inference methods. We adopt the *sequential tree-reweighted message passing* (TRW-S) proposed in [Kolmogorov 2006] (see also section 2.2.3), since it offers a good compromise between the quality of the obtained minimum, the ability to model complex interactions between the nodes and reasonable computational complexity.

3.4.2 Results

We show the results on two sequences with rigid objects and three sequences with deformable objects. The test sequences have been degraded (severe noise has been added to some of them), while at the same time present varying complexity with respect to the objects and background visual properties, varying degrees of occlusions, as well as static and moving observers.

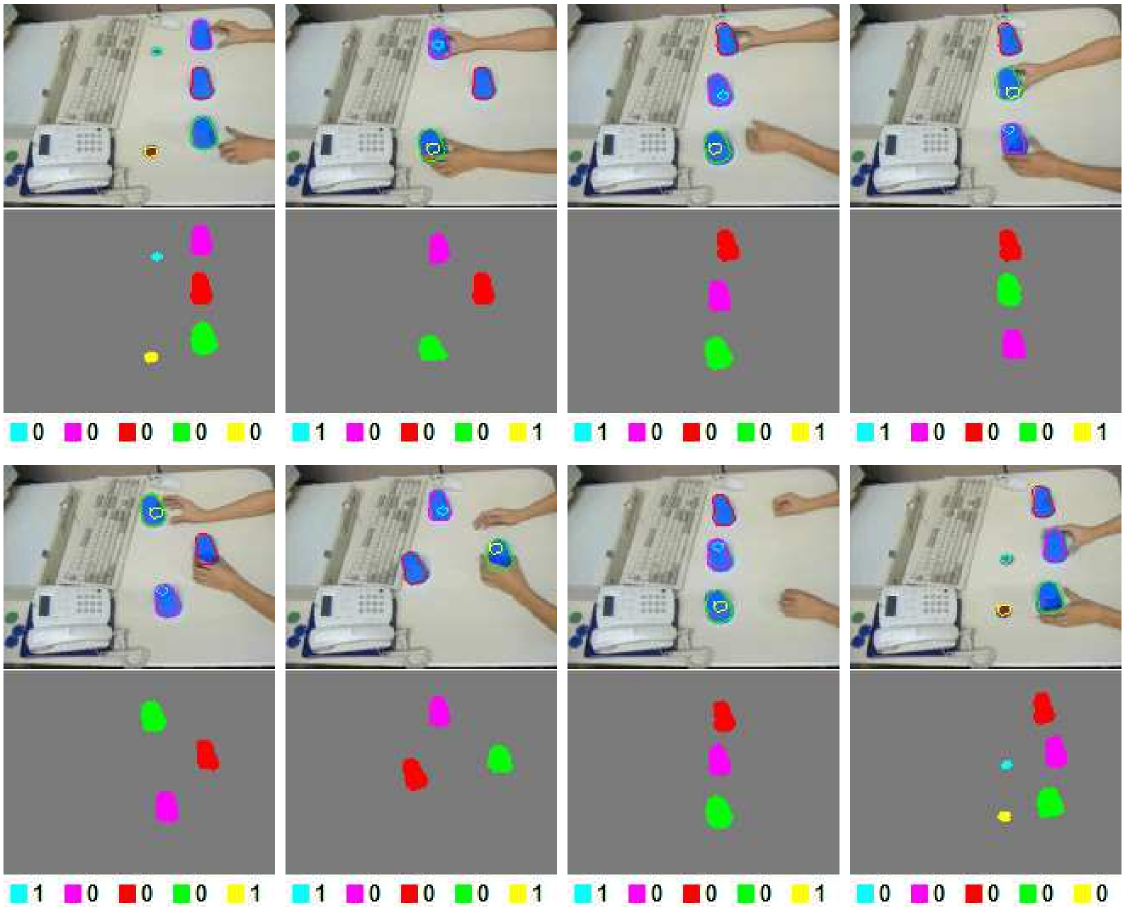


Figure 3.5: Experimental Results for *Shell Game Sequence*.

Box Sequence

In the original sequence, two boxes move such that significant occlusions (including complete occlusions) occur between them. Our algorithm has successfully tracked the objects, segmented the image, and estimated the depths of the objects. Furthermore, in order to test the robustness to noise, we independently added Gaussian white noise of mean 0 and variance 0.8 (the range of RGB value is $[0, 1]^3$) to each frame, and then tested our method. Our method still has performed very well despite the presence of severe noise and occlusions. Fig. 3.4 shows the obtained results on this very degraded video.

Shell Game Sequence

In order to test the robustness with respect to temporally and spatially significant occlusions, we have considered *Shell Game* sequence [Huang & Essa 2005]. In this video, there are three identical cups facing downwards and two chips of different colors. The opera-

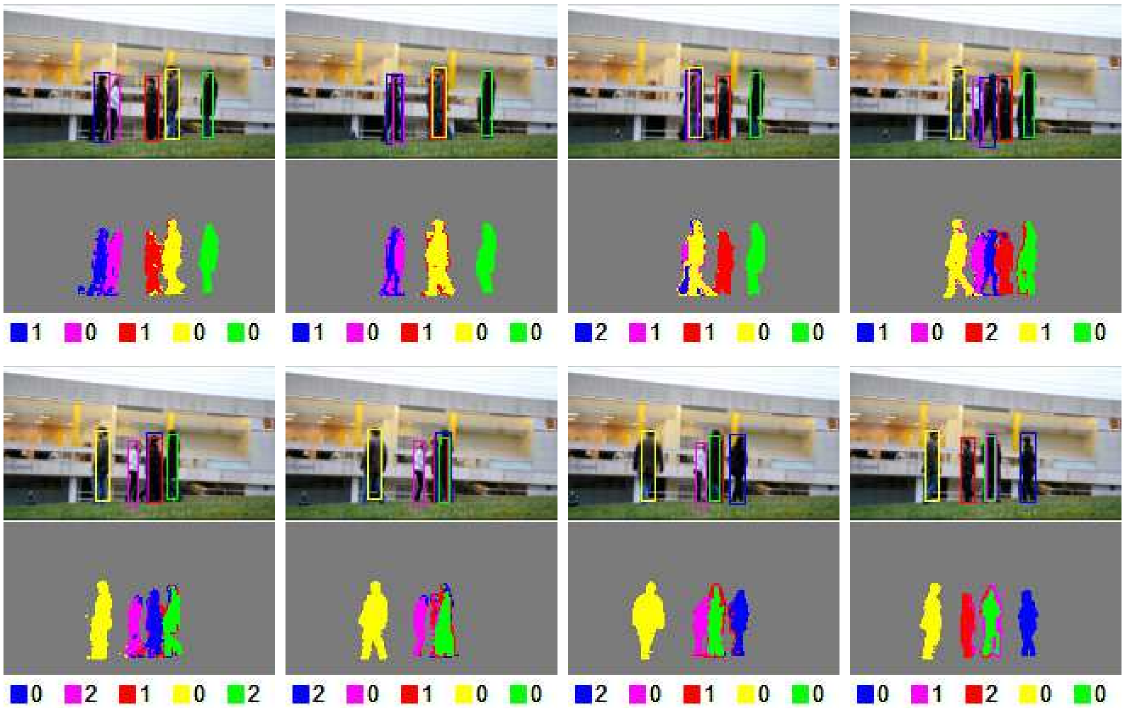


Figure 3.6: Experimental Results for *Pedestrian Sequence 1*.

tor begins the game by placing two cups such that each cup covers one of the two chips, then he/she quickly shuffles the three cups around and finally uncovering the chips. While being occluded, each chip keeps sliding with the cup that covers it. This video is quite challenging mainly due to the long-term complete occlusions of the two occluded chips (Fig. 3.5).

Note that we previously assumed that the background was always behind all the objects. However, one can also imagine floating background, *i.e.*, those objects which are not to be tracked but may occlude the tracked objects (*e.g.* the hands in the video). Such a floating background can also be modeled as a layer in our model. In our experiments, we dealt with this by adding another possible depth "-1" for the background (*i.e.*, add $(0, -1)$ into $\mathcal{X}_i$ $(i \in \mathcal{V}_p)$) and giving a prior penalty to the case where a pixel is labeled as "background" and has depth "-1".

Pedestrian Sequences

Sever occlusions have also been considered in a real setting, with deformable objects, image noise, changes of illumination and moving camera. We have considered three sequences: (i) the first one consists of a static background with five people, severe occlusions between the objects and the maximum level of occlusions being three (Fig. 3.6); (ii) the

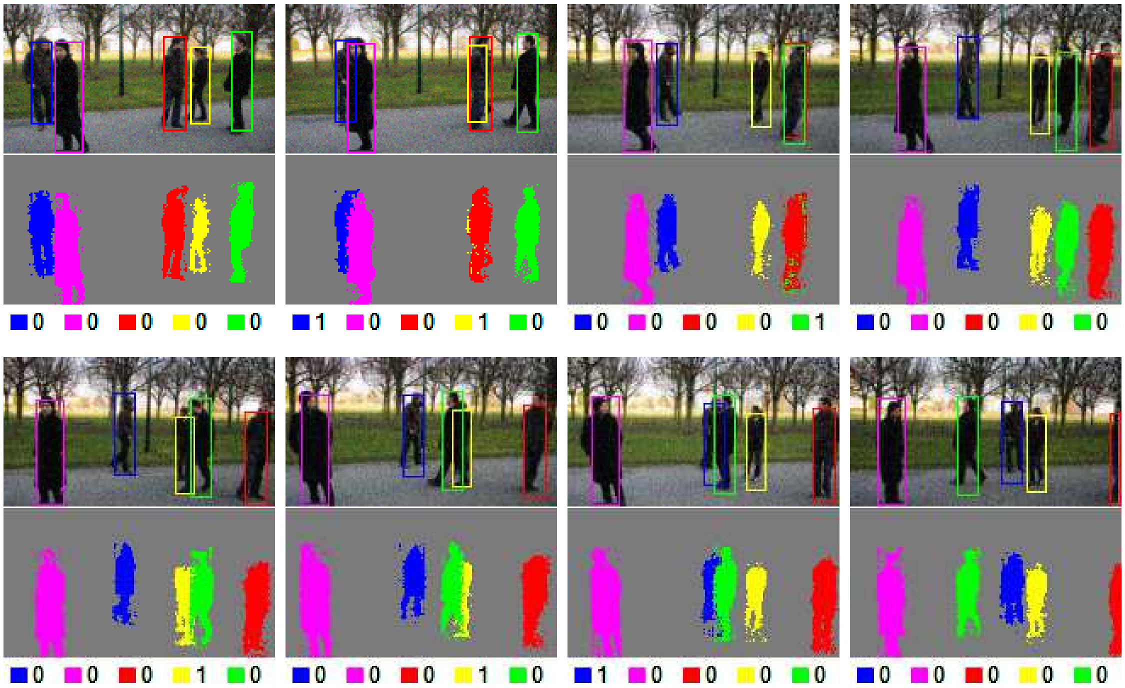


Figure 3.7: Experimental Results for *Pedestrian Sequence 2*.

second one consists of a moving background with five people, severe noise and changes of illumination (Fig. 3.7); (iii) the last one consists of a moving background with four people and significant changes in texture (Fig. 3.8).

For these pedestrian sequences, a rectangle is used to model the shape of a person. Since, in the shape prior, the torso is more reliable than the limbs due to limb motions, we manually set an area inside the shape model (*i.e.*, including the majority of the torso), and it has the same motion as the shape model. When computing $\psi_{ki}^{(3)}(x_k, x_i)$ using Eq. 3.16, if pixel i is inside this area with the configuration θ_k, we multiply $\psi_{ki}^{(3)}(x_k, x_i)$ by a factor 10 to increase the confidence to this area, and otherwise we divide $\psi_{ki}^{(3)}(x_k, x_i)$ by a factor 10 to decrease the confidence.

For all these test sequences, our algorithm has successfully segmented, tracked and ordered by depth all the objects. We believe that the robustness is due to the strategy of coupling segmentation, tracking and depth ordering in a unified single-shot optimization formulation and also due to the high quality of the optimum provided by the TRW-S algorithm. The main limitation of the method is the computational complexity. Running times vary from a few seconds to several minutes per frame. It is shown that, with presence of occlusions in the observed image, TRW-S requires much more iterations to converge than the cases without occlusions. The theoretical justification of added complexity in the presence of occlusions and development of specific optimization algorithms for solving such

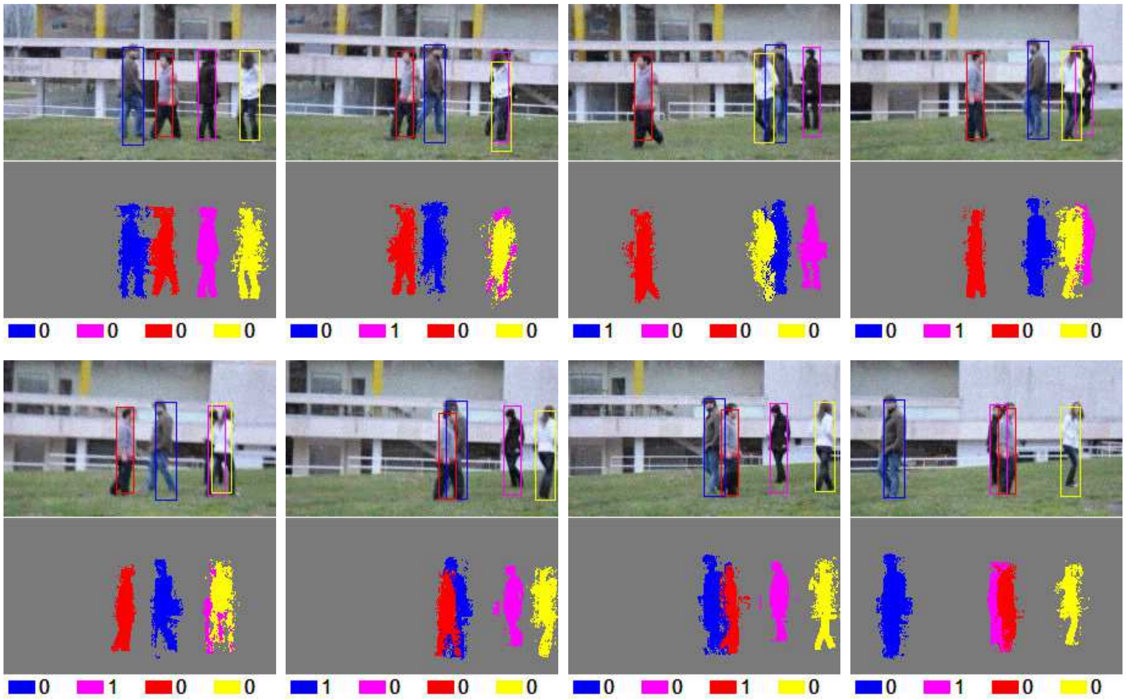


Figure 3.8: Experimental Results for *Pedestrian Sequence 3*.

special MRFs more efficiently are interesting problems to be explored in the future works.

3.4.3 Discussion

Algorithm Acceleration

In the general MRF formulation (section 3.3), all the objects are connected with all the pixels. However, in the tracking scenario, such pixel-object connections can be significantly relaxed by considering the temporal consistency on the motion configuration, as in general the object motion is bounded in a finite speed. Based on this observation, we propose an approach to simplifying the MRF model in section 3.3. For an object k $(k \in \mathcal{V}_o)$, once we get the estimation of its motion parameters $\theta_k^{\text{opt},(t-1)}$ and predict the parameters $\hat{\theta}_k^{(t)}$ for instant t, we compute the distance function $\mathcal{M}_k(\hat{\theta}_k^{(t)})$. Using this distance function, we prune the connections between the object k and those pixels i with $dist(i, \mathcal{M}_k(\hat{\theta}_k^{(t)})) > b$, where b is a tolerance coefficient. And for these pixels, the label k is excluded from their configuration spaces. In this way, the complexity of the MRF model is reduced with respect to both the topology and the space of latent variables. In the experiments, we observed that the algorithm can be sped up by more than 15 times on average (with $b = 20$).

Introducing Interactions between Pixels

As we said previously, we can also introduce interactions/constraints on the labels of the pixel nodes through conventional 4-neighborhood or 8-neighborhood systems. To this end, we add the edges between those neighbor pixels into the edge set $\mathcal{E}$. Thus, we can smooth the segmentation result using Potts model [Potts 1952] by defining the corresponding potential as:

$$\psi(x_i, x_j) = \begin{cases} \eta \; (\eta > 0) & \text{if } l_i \neq l_j \\ 0 & \text{if } l_i = l_j \end{cases} \quad (i, j \in \mathcal{V}_p, (i,j) \in \mathcal{E}) \tag{3.20}$$

which favors neighbor pixels having the same label. We can also define other forms of potentials (*e.g.*, by considering the contrast). We have tested the cases both with and without this smoothness term. It is shown that the inclusion of this term does not improve the tracking performance but can smooth and improve the segmentation to some extent. However, the running-time significantly increases with the use of this term and the choice of η complicates the parameter setting.

3.5 Conclusion

In this chapter, we have proposed a novel single-shot optimization approach for segmentation, depth ordering and tracking with occlusion handling. Our approach is based on our joint layered image modeling, where a distributed way has been introduced to deal with visibility satisfaction where individual pixel modeling contributes to the depth ordering of objects through local condition preservation constraints. The above constraints are expressed as cost terms in an MRF and are integrated with image support towards scene understanding. To the best of our knowledge, this is the first approach that combines low-level image support with high-level object representation along with rigorous occlusion handling in a single modular MRF where image data terms as well as priors can be easily replaced with more advanced models. Promising experimental results demonstrate the potential of the method. However, only weak shape priors such as rectangles were considered. Towards introducing richer high-level shape prior knowledge into grouping problems, we have studied the problem of non-rigid 3D surface matching, which will be presented in the next chapter (chapter 4).

Indeed, jointly modeling high-level knowledge about the scene and low-level image evidence using a principled formulation can highly improve the performance and robustness of a method for the inference about the scene and the image, and graphical models provides a powerful tool to achieve such a modeling. Besides the joint segmentation, tracking and depth ordering as presented here, the same insight has also been employed in another work [Panagopoulos *et al.* 2010, Panagopoulos *et al.* 2011] which I have participated into, where we aim to jointly recover the illumination environment and an estimate of the cast shadows in a scene from a single image, given coarse 3D geometry. For this objective, we proposed in [Panagopoulos *et al.* 2011] a higher-order MRF illumination model to jointly model the illumination environment and the intensity values of pixels, where the consistency between the illumination condition and the intensity value of pixels are encoded within higher-order clique potentials and all the latent variables are simultaneously inferred through the minimization of the energy of the MRF. Despite the fact that the geometry used in the experiments consists of bounding boxes or a common rough 3D model for a whole class of objects, our MRF illumination model still have achieved high-quality estimation results on various datasets.

Chapter 4

Higher-order Non-rigid 3D Surface Matching

In this chapter, we aim at developing a robust algorithm for non-rigid 3D surface matching. To this end, we propose a higher-order graph-based formulation, where singleton terms encode geometric and appearance similarities (*e.g.*, curvature and texture), while higher-order terms capture intrinsic deformation errors. The pseudo-boolean representation of the objective function involved in such a formulation is optimized using a dual-decomposition-based method to achieve optimal correspondences between two surfaces. Furthermore, an efficient two-stage optimization approach is introduced towards achieving dense surface matching. Our method has been validated through a series of experiments, which demonstrate its accuracy and efficiency, notably in challenging cases of large and/or non-isometric deformations, or meshes that are partially occluded.

4.1 Introduction

Surface matching (also known as *surface registration* or *surface alignment*), whose objective is to determine meaningful correspondences between two or more surfaces, is a fundamental problem in computer vision, computer graphics and medical imaging for numerous important applications such as 3D shape retrieval, deformation transfer, object recognition, facial expression recognition, statistical shape modeling and shape change detection (see [Campbell & Flynn 2001, van Kaick *et al.* 2010]). Nowadays, surface matching has become even more important due to the rapid development of 3D acquisition techniques (*e.g.*, [Zhang *et al.* 2004, Wang *et al.* 2005, Hernández *et al.* 2007, Shaji *et al.* 2010, Kinect 2010]) and the desire for building various attractive applications on these 3D data. Despite a large amount of literature on surface matching (see [van Kaick *et al.* 2010] for a survey of methods), it remains a very challenging problem, particularly when the surfaces undergo large, non-rigid deformations and are subject to a high level of noise. In order to handle surface matching problems under such difficult situations, it is usually necessary to take into account both local feature similarities and global deformation constraints. While local structures are somewhat straightforward to handle, the consideration of global structures imposes a major challenge for surface matching. Another difficulty lies in the inherent complexity of the problem, *i.e.*, the matching problem is a combinatorial problem and the number of possible matching configurations is $N!$ for the case of bijective matching (where N denotes the number of points on each surface), which will become even larger if partial matching is allowed.

In order to impose global deformation constraints for the surface matching problem, many existing works are based on certain *rigidity* assumptions on the deformation of the surface and impose rigidity as a global regularization when searching for correspondences. Assuming that two surfaces only differ by a global rigid deformation (*i.e.*, rotation and translation), the iterative closest points (ICP) [Besl & McKay 1992] method and its variants [Rusinkiewicz & Levoy 2001] have been successfully applied for near-rigid surface registration with various extensions (*e.g.*, [Hahnel *et al.* 2003, Brown & Rusinkiewicz 2007]). In such a context, the global distortion is defined on the correspondence configurations of all the points (referred to as *matching configuration*) that are determined by the configuration of the global pose. To minimize such a distortion, the ICP algorithm alternates between establishing correspondences given the rigid transformation and estimating the rigid transformation given the correspondences. Obviously, such a scheme easily gets stuck in local minima and thus requires that the initial poses of the two surfaces are close enough to get a satisfactory matching result. Moreover, global rigidity does not take into account bendable shapes (*e.g.*, garments or rubber bands) and thus makes it difficult to deal with surfaces undergoing large non-rigid deformations. To deal with this, the notion of *local rigidity* has been proposed to model non-rigid deformations, by assuming that the deformation between two local neighborhoods of each correspondence is rigid (*e.g.*, [Huang *et al.* 2008]). Similar to the ICP algorithm, an alternating optimization scheme is usually required to optimize the objective function, which severely limits the quality of solution, especially when registering two surfaces with large deformations. Also, considering local rigidity in the 3D space is challenging when the deformation between two shapes are large, due to the lack of efficient optimization techniques for such a large search space.

For the non-rigid surface matching problem, most of the methods in the literature are based on a common assumption that the undergoing deformation between two surfaces is *isometric*, which means that the lengths of any infinitesimal vectors between a pair of corresponding points is preserved. For a surface undergoing isometric deformation, the geodesic distance between any pair of points on the first surface is the same as the geodesic distance between their corresponding points on the second surface. Mathematically, let $d_{\mathcal{P}_1}(p,q)$ denote the geodesic distance between two points p and q on the surface $\mathcal{P}_1$, we have the following definition:

Definition 2. *A map $f : \mathcal{P}_1 \rightarrow \mathcal{P}_2$ is isometric if and only if the following condition holds:*

$$d_{\mathcal{P}_1}(p,q) = d_{\mathcal{P}_2}(f(p), f(q)), \ \ \forall p,q \in \mathcal{P}_1.$$

The isometric assumption is a good approximation to many real-world deformations. For example, the deformation of a cloth is usually isometric and the deformation of face is nearly isometric [Bronstein *et al.* 2007]. Compared with those methods based on rigidity or local rigidity assumption, approaches based on isometric assumption or geodesic distances between pairs of points on the surface exhibit better performance when dealing with large deformations. [Elad & Kimmel 2001] proposed to use multidimensional scaling (MDS) to represent shapes in a low-dimensional Euclidean space such that the geodesic distances between a pair of points in the original space are closely approximated by Euclidean distances in the embedding space. Then the surface matching can be done by comparing them as rigid objects in such an embedding space. The use of an intermediate embedding space was eliminated in [Mémoli & Sapiro 2005] by using the Gromov-Hausdorff formalism [Gromov 1981]. [Bronstein *et al.* 2006] proposed an MDS-like algorithm referred to as generalized MDS (GMDS) for the computation of the Gromov-Hausdorff distance and deformation invariant correspondence between shapes. This framework was extended in [Bronstein *et al.* 2010] using diffusion geometry instead of the geodesic one. Such methods usually inherit embedding errors and do not consider extrinsic information when establishing correspondences. Another approach without explicit embedding is to formulate surface matching as an MRF optimization problem (*e.g.*, [Anguelov *et al.* 2004]), where pairwise potentials between neighbor points are defined based on the deviation of geodesic distance and the *loopy belief propagation* algorithm is used for the MRF optimization. Nevertheless, the deviation of geodesic distance is still a local measurement of the quality of surface matching in the sense that it does not take into account the information about the matching of other points (due to lack of a "global" view on the matching of the whole surface). As a result, those methods where only geodesics are considered may suffer from certain "geodesic" ambiguities and are not robust enough in cases where the data are corrupted by noise.

There is also a family of approaches for matching surfaces with large deformations based on the conformal mapping, such as the works in [Wang *et al.* 2007, Zeng *et al.* 2008, Zhang & Hebert 1999]. An important property of conformal mapping is that if two surfaces are isometrically deformed from one to the other, their correspondences only dif-

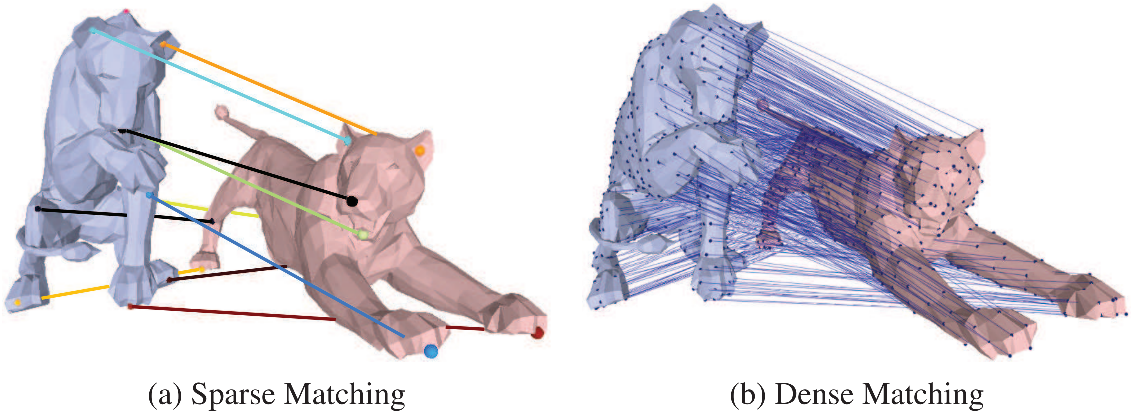

(a) Sparse Matching (b) Dense Matching

Figure 4.1: Matching Result between Two Surfaces. This result between two surfaces undergoing a large non-rigid deformation demonstrate the performance of our approach in establishing both the sparse (a) and dense (b) correspondences.

fer by a Möbius transformation in their conformal parametrization (also known as *uniformization*) domains. Hence, once such a transformation is recovered, one-to-one correspondences between the two surfaces can be established, giving us a global transformation between two surfaces. Based on such a global property of conformal mapping, [Wang *et al.* 2007, Zeng *et al.* 2008, Zhang & Hebert 1999] established dense correspondences between two surfaces by specifying a few initial feature correspondences. As a result, the performance of these approaches relies heavily on the accuracy of the selection of the initial correspondence points. To remedy this, [Lipman & Funkhouser 2009] proposed to find sparse correspondences between two surfaces based on a voting scheme. Since every three correspondences determine a unique Möbius transformation between the uniformization domain of the two surfaces, they also determine a correspondence mapping between two surfaces. Hence, for each possible triplet of correspondences, one can define a measure of the plausibility (or metric) of such correspondences by matching among all the other points on the whole surface using the Möbius transformation recovered from the triplet of correspondences. Despite promising performance of such a voting scheme, a main drawback is that there is no guarantee on the quality of the final results, since the voting scheme does not optimize a concrete objective function. Also, only intrinsic deformation information was considered in [Lipman & Funkhouser 2009], while a principled integration with other cues such as extrinsic similarity information would be difficult to be done in such a voting scheme. However, the proposed metric for measuring the quality of any triplet correspondences provides us a way to measure the global distortion locally by considering a triplet of correspondences.

Graph Matching

Many computer vision and pattern recognition problems can be formulated as a graph matching problem [Conte *et al.* 2004]. Mathematically, the *bipartite graph matching* problem is defined on a bipartite graph $\mathcal{G} = (\mathcal{U}, \mathcal{V}; \mathcal{E})$ where $\mathcal{U}$ and $\mathcal{V}$ denote two disjoint node sets and $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{V}$ denotes an edge set. A matching $\mathcal{M}$ is a subset of the edge set $\mathcal{E}$ such that every node of $\mathcal{G}$ appears in at most one edge in the matching $\mathcal{M}$ [Lovasz & Plummer 1986]. A matching is called *perfect* when $|\mathcal{U}| = |\mathcal{V}|$ and every node of $\mathcal{G}$ coincides with one and only one edge of the matching $\mathcal{M}$. [Hall 1935] proved the *marriage theorem* which gives a necessary and sufficient condition for the existence of a perfect matching. Moreover, a perfect matching in bipartite graphs can be found in $O(|\mathcal{E}|\sqrt{|\mathcal{U} \cup \mathcal{V}|})$ time using the *Hopcroft-Karp algorithm* [Hopcroft & Karp 1973].

If we assign a weight (or cost) to each correspondence (*i.e.*, each edge in $\mathcal{E}$) and aim to find the optimal matching whose sum of weights are minimal (or maximal), then such a problem is referred to as *minimum weight matching* (or *maximum weight matching*). The cost can be defined on each correspondence, a pair of correspondences and/or multiple correspondences to constrain the configuration of matching. There are various matching problems referred to as *linear assignment problems*, *quadratic assignment problems* and *multi-index assignment problems*, according to the maximal number of correspondences that are assigned the cost.

In the linear assignment problem, cost functions are only defined on individual correspondences (referred to as *singleton potentials*). The *Hungarian algorithm* was proposed in [Kuhn 1955] to find a perfect matching with minimum cost, which is the genesis of the network flow based algorithm that later gained widespread popularity in the combinatorial optimization community. The computational complexity of the original algorithm of [Kuhn 1955] is $O(n^4)$, which was later reduced to $O(n^3)$ in [Dinic & Kronrod 1969].

Besides singleton potentials, the quadratic assignment problem (corresponding to *pairwise graph matching*) considers cost functions defined on pairs of correspondences (corresponding to *pairwise potentials*) as well. Quadratic assignment problems provides a powerful tool for modeling numerous real-world applications, due to the consideration of interactions between a pair of correspondences. In computer vision problems, the matching cost is often used to measure the dissimilarity of two graphs, where pairwise potentials can model soft contextual constraints (similar to pairwise MRFs). It has been previously used to deal with various vision problems such as shape matching and object recognition (*e.g.*, [Belongie *et al.* 2002, Berg *et al.* 2005]), the matching of feature points (*e.g.*, [Leordeanu & Hebert 2005, Torresani *et al.* 2008]) and character recognition (*e.g.*, [Rocha & Pavlidis 1994, Lee & Liu 1999]). However, solving a quadratic assignment problem is an NP-hard problem [Sahni & Gonzalez 1976]. Numerous methods have been proposed to deal with such a problem. One can cite for example branch-and-bounds approaches (*e.g.*, [Tsai & Fu 1979, Cordella *et al.* 1996, Cordella *et al.* 2001]), spectral relaxation methods (*e.g.*, [Umeyama 1988, Carcassoni & Hancock 2003, Caelli & Kosinov 2004, Leordeanu & Hebert 2005, Cour *et al.* 2007]), methods based on continuous relaxation (*e.g.*, [Gold & Rangarajan 1996, Torr 2003, Schellewald & Schnorr 2005]), *etc.* As pointed out by [Torresani *et al.* 2008], most of these methods have no optimality guarantee. In such a context, [Torresani *et al.* 2008] proposed a novel pairwise graph-matching algorithm based on the well-known dual-decomposition optimization framework (see section 2.2.3), which provides optimality guarantee and exhibits very promising matching performance.

Multi-index assignment problems (corresponding to *high-order graph matching*) consider higher-order interactions between three or more correspondences. There are various works that explored higher-order similarity measures to improve the matching accuracy, resulting in high-order graph matching problems (*e.g.*, [Zass & Shashua 2008, Duchenne *et al.* 2009, Chertok & Keller 2010]). Obviously, the optimization of such problems is even harder than that of quadratic assignment problems in general. In order to solve the high-order graph matching formulations, [Zass & Shashua 2008] proposed a probabilistic approach and [Duchenne *et al.* 2009, Chertok & Keller 2010] developed spectral relaxation methods based on the optimization algorithms for pairwise counterparts [Leordeanu & Hebert 2005], from which one can expect their optimality properties would be similar to those for pairwise counterparts. In such a context, we are motivated to use the same insight as [Torresani *et al.* 2008] and recent order-reduction techniques (*e.g.*, [Ishikawa 2009]) to deal with high-order graph-matching problems.

4.1.1 Our Approach

Our goal is to robustly establish the correspondences between two non-rigid surfaces undergoing large (near-isometric) deformations and possibly partial matching, without requiring correspondence initialization and alternating search.

In order to achieve a robust matching, it is desirable to consider the structure of the surfaces at both local and global levels [van Kaick *et al.* 2010] and to encode the distortion at both levels within a single formulation that is able to be solved efficiently. A graph-based formulation provides a sound mathematical tool that allows to define such a matching cost and perform efficient optimization. However, defining a robust global distortion in a graph-based formulation imposes a challenge, although local matching costs can be defined conveniently by measuring the similarity or distortion between local structures and encoded within for example singleton potentials. Fortunately, conformal mapping theory provides an efficient way to measure the similarity of global structures between two surfaces. Based on the fact that three correspondences can determine a mapping between two surfaces, we can then measure the global distortion induced by such a triplet of correspondences using the deviation of such a mapping from isometry, which can be done efficiently using the metric proposed in [Lipman & Funkhouser 2009]. This observation motivates us to define third-order potential functions to encode the global distortion that is implied by any possible triplet of correspondences.

In summary, we propose a novel approach to robustly establish correspondences between two surfaces via a high-order graph matching formulation. More specifically, we consider multiple measurements (*e.g.*, curvature, texture) to capture the appearance and geometric similarity between local structures and third-order interactions to model the distortion of global structures (*i.e.*, intrinsic deformation errors) between a triplet of correspondences. All these measurements are integrated within a higher-order graph matching framework which is represented using a pseudo-boolean function [Boros & Hammer 2002]. In order to optimize such a higher-order function, we reduce the third-order potentials to quadratic terms [Ishikawa 2009] and obtain a near optimal solution based on the dual-decomposition technique [Bertsekas 1999, Komodakis *et al.* 2007a]. Last but not least, towards dense surface matching, a hierarchical algorithm is introduced to constrain the search space through candidate selection and local graph matching. The whole method is able to establish dense matching between surfaces undergoing large non-rigid (near-isometric) deformations, partial matching and even inconsistent boundaries and scales.

4.1.2 Outline of the Chapter

The reminder of this chapter is organized as follows. We present in section 4.2 the proposed high-order graph-based formulation for surface matching problem. Then in section 4.3, we present the two-stage hierarchical surface matching framework for dense surface matching. Experimental validation are presented in section 4.4. Finally, we conclude this chapter in section 4.5.

4.2 Higher-order Surface Matching Formulation

In this section, we formulate surface matching as a high-order graph matching problem, where an objective function is defined based on various measurements of geometric/appearance similarities and intrinsic deformation errors and then is to be optimized to achieve matching results.

4.2.1 Pseudo-boolean Formulation

Let us denote by $\mathcal{P}_1$ and $\mathcal{P}_2$ the set of points from two surfaces S_1 and S_2, respectively. Then, $\mathcal{A} \triangleq \mathcal{P}_1 \times \mathcal{P}_2$ denotes the set of possible correspondences (also referred to as *assignments*). Similar to [Torresani *et al.* 2008], we define a boolean indicator variable for each potential correspondence $a = (i, j) \in \mathcal{A}$ to characterize its state (*active*[1] or *inactive*):

$$x_a = \begin{cases} 1 & \text{if } a = (i, j) \in \mathcal{A} \text{ is active} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Hence, the joint variable $\mathbf{x} = (x_a)_{a \in \mathcal{A}}$ denotes the activation states of all the correspondences (*i.e.*, *matching configuration*).

A basic constraint imposed on the matching configuration is that each point in $\mathcal{P}_1$ is mapped to at most one point in $\mathcal{P}_2$, while for each point in $\mathcal{P}_2$ there is at most one point in $\mathcal{P}_1$ mapping to it. Note that a point is allowed to have no active correspondence in order to deal with partial matching. Under such a constraint, we can define the feasible solution space $\mathcal{X}$ of the matching configuration $\mathbf{x}$ as follows:

$$\mathcal{X} = \{\mathbf{x} \in \{0, 1\}^{|\mathcal{A}|} | \sum_{i \in \mathcal{P}_1} x_{i,j} \leq 1, \sum_{j \in \mathcal{P}_2} x_{i,j} \leq 1, \forall\, i \in \mathcal{P}_1 \text{ and } \forall\, j \in \mathcal{P}_2\} \tag{4.2}$$

As we mentioned in section 2.1.3, higher-order models allow to naturally model certain measures that cannot be encoded using pairwise ones such as second-order derivative and scale invariant measures in Euclidean space. In this work, we propose a third-order graph matching formulation to deal with the surface matching problem, where an energy function $E(\mathbf{x})$ consisting of singleton, pairwise and third-order terms is defined and minimized over $\mathcal{X}$ to achieve the optimal matching configuration, *i.e.*,

$$\mathbf{x}^{\text{opt}} = \arg \min_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}) \tag{4.3}$$

The energy function $E(\mathbf{x})$ has the following form:

$$E(\mathbf{x}) = \sum_{a \in \mathcal{A}} \theta_a x_a + \sum_{(a,b) \in \mathcal{A} \times \mathcal{A}} \theta_{ab} x_a x_b + \sum_{(a,b,c) \in \mathcal{A} \times \mathcal{A} \times \mathcal{A}} \theta_{abc} x_a x_b x_c \tag{4.4}$$

where θ_a is the singleton matching cost for each active correspondence $a \in \mathcal{A}$, θ_{ab} for a pair of active correspondences $(a, b) \in \mathcal{A} \times \mathcal{A}$, and θ_{abc} for a triplet of active correspondences $(a, b, c) \in \mathcal{A} \times \mathcal{A} \times \mathcal{A}$. In fact, the matching constraint in Eq. 4.2 can be reduced to pairwise terms in the energy function by using the following equivalence:

$$\forall\, i \in \mathcal{P}_1, \sum_{j \in \mathcal{P}_2} x_{i,j} \leq 1 \text{ iff } \min_{\mathbf{x}} \sum_{j', j'' \in \mathcal{P}_2, j' \neq j''} \theta^{\infty} x_{i,j'} x_{i,j''} = 0 \tag{4.5}$$

where θ^{∞} denotes a sufficiently large number. Let us use $\mathcal{A}^{\mathcal{C}}$ to denote the set of pairs that encodes the matching constraints for all the correspondences. Thus, the high-order

[1] A potential correspondence is *active* means that it is included in the matching.

matching problem can be formulated as the following pseudo-boolean optimization problem [Boros & Hammer 2002] as follows:

$$\min_{\mathbf{x}\in\{0,1\}^{|\mathcal{A}|}} \{E(\mathbf{x}) = \sum_{a\in\mathcal{A}} \theta_a x_a + \sum_{(a,b)\in\mathcal{A}\times\mathcal{A}} \theta_{ab} x_a x_b + \sum_{(a,b)\in\mathcal{A}^{\mathcal{C}}} \theta^{\infty} x_a x_b + \sum_{(a,b,c)\in\mathcal{A}\times\mathcal{A}\times\mathcal{A}} \theta_{abc} x_a x_b x_c\} \tag{4.6}$$

The above formulation is general and can capture different matching scenarios, including partial matching, by properly defining the potentials.

Due to the positive weight θ^{∞} that encodes the matching constraint, the energy function 4.6 is non-submodular [Freedman & Drineas 2005] and the minimization of such an energy is an NP-hard problem in general [Boros & Hammer 2002]. An advantage of the pseudo-boolean formulation is that any high-order terms can be reduced into a quadratic term [Boros & Hammer 2002], which can then be solved by existing efficient optimization algorithms such as QPBO techniques [Boros *et al.* 1991, Kolmogorov & Rother 2007, Boros *et al.* 2006, Rother *et al.* 2007]. Such a reduction can be done efficiently using for example the reduction method recently proposed in [Ishikawa 2009].

4.2.2 Definition of Potential Functions

In order to consider multiple sources of similarity measurements, the potential functions in Eq. 4.4 are defined such that both local features and global deformation information contribute to the objective function. In this work, only singleton and third-order terms are considered for simplification, where the singleton terms are used to measure the dissimilarity of local structures while the third-order terms take the distortion of global structure into account. Note that pairwise potentials can also be considered in this general formulation to integrate more geometric information towards improving the matching performance. For example, we can encode geodesic [Mémoli & Sapiro 2005, Bronstein *et al.* 2006], diffusion metrics [Bronstein *et al.* 2010] and commute time metrics [Qiu & Hancock 2007] on the surface within pairwise potentials.

Singleton Potentials

Singleton potentials encode geometric and/or photometric compatibility between the local structures of an active correspondence, as in [Thorstensen & Keriven 2009]. For simplicity, we use the Gaussian curvature $\text{curv}(i)$ at point i as geometric descriptor, which is invariant to isometric transformation [do Carmo 1976], and the texture value $\text{tex}(i)$ at point i as photometric descriptor if the texture information is available. Then, the singleton potential for a correspondence (i, j) is defined as follows to favor correspondences having similar local structures:

$$\theta_{i,j} = (\text{curv}(i) - \text{curv}(j))^2 + \lambda_0(\text{tex}(i) - \text{tex}(j))^2 \tag{4.7}$$

where λ_0 is a positive coefficient that balances the contribution between the curvature and the texture information. Similarly, other features can also be considered within such potentials such as spin-image [Johnson 1997], multiscale heat kernel signatures [Sun *et al.* 2009, Ovsjanikov *et al.* 2009, Bronstein & Kokkinos 2010], eigenfunctions of the Laplace-Beltrami operator [Rustamov 2007, Mateus *et al.* 2008, Hu & Hua 2009, Dubrovina & Kimmel 2010] and local photometric properties [Zaharescu *et al.* 2009, Thorstensen & Keriven 2009].

Higher-order Potentials

High-order potentials encode the distortion of global structures for any triplet of correspondences as well as the consistency of extrinsic orientations.

According to the uniformization theorem [Farkas & Kra 2004], any 3D surface can be flattened conformally to a canonical 2D domain. Within such a mapping each feature point p has a parametric coordinate in the complex plane $z_p \in \hat{\mathbb{C}}$. If the undergoing deformation between two surfaces is isometric, then the mapping between their parameterizations in the 2D domain is a Möbius transformation, which can be uniquely determined by considering three pairs of corresponding points on the surfaces (a triplet of points from each surface). Inspired by [Lipman & Funkhouser 2009], we compute the intrinsic deformation error based on the Möbius transformation as the distortion of global structures induced by two corresponding triplets.

Given two surfaces, S_1 and S_2, for any two triplets, $(p_i^1, p_j^1, p_k^1) \in S_1$ and $(p_i^2, p_j^2, p_k^2) \in S_2$, we first recover the associated Möbius transformation $m^1(z)$ and $m^2(z)$ that map each triplet to a constant configuration $(e^{i\frac{2\pi}{3}}, e^{i\frac{4\pi}{3}}, e^{i2\pi})$. This transformation essentially equips each point in the sets P_1 and P_2 with coordinates in $\hat{\mathbb{C}}$. Let us denote the new coordinate for each point p as $z(p) \in \hat{\mathbb{C}}$. Similar to [Lipman & Funkhouser 2009], we establish correspondences between the two sets P_1 and P_2 by searching the mutually closest point correspondences set under the new coordinates, denoted as:

$$\begin{aligned}\mathcal{M}_{ijk} = \{&(p_1, p_2) | p_1 \in S_1, p_2 \in S_2, \text{such that:}\\ &\forall\, p_2' \in S_2 \setminus \{p_2\}, |z(p_1) - z(p_2)| < |z(p_1) - z(p_2')|,\\ &\forall\, p_1' \in S_1 \setminus \{p_1\}, |z(p_1) - z(p_2)| < |z(p_1') - z(p_2)|\}\end{aligned} \tag{4.8}$$

and define the deformation error as

$$E_{ijk} = \sum_{(p_1,p_2)\in\mathcal{M}_{ijk}} |z(p_1) - z(p_2)|^2 \tag{4.9}$$

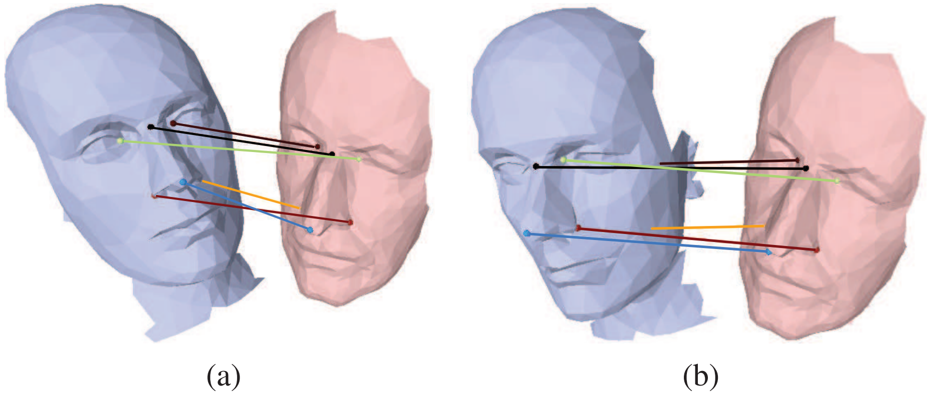


Figure 4.2: Example of Ambiguity due to Intrinsic Symmetry. This figure shows the ambiguity by considering only the intrinsic embedding information. The matching scores in (a) and (b) are the same from Eq. 4.9 based on the Möbius transformation, since the distances between the matching features are identical. However, such ambiguity can be avoided by adding the extrinsic similarity information (*e.g.*, normal and curvature).

Then we define the Möbius matching potential as follows,

$$\theta_{ijk}^{\text{Möbius}} = \begin{cases} \frac{E_{ijk}}{|\mathcal{M}_{ijk}|^2} - 1 & \text{if } \frac{E_{ijk}}{|\mathcal{M}_{ijk}|} < \delta \\ 1/|\mathcal{M}_{ijk}| & \text{otherwise} \end{cases} \tag{4.10}$$

Here δ is a lower bound value to single out unlikely correspondences (in our experiment $\delta = 0.1$). Without it the minimization problem of Eq. 4.4 would encourage as many correspondences as possible even when some of them do not match. Intuitively, if there were more matching pairs and the distances between those matching pairs were smaller, the potential would be lower.

However, considering the Möbius energy alone can introduce a certain ambiguity, since it encodes only isometric information (an example is shown in Fig. 4.2). In order to eliminate such an ambiguity, we consider the Gaussian map of the surface. The Gaussian map is defined as the mapping of the normal at each point on the surface to the unit sphere [do Carmo 1976]. The Gaussian map captures the extrinsic geometric information of the surface. In order to avoid ambiguities in orientation, the orientation of the Gaussian maps is considered for each of the triplets. Two triplets have the same orientation if and only if the determinant of their normals have the same sign. Therefore, we define another higher-order term as follows:

$$\theta_{ijk}^{\text{Gaussian}} = \begin{cases} 0 & \text{if } \det(\mathbf{n}_i^1, \mathbf{n}_j^1, \mathbf{n}_k^1) \cdot \det(\mathbf{n}_i^2, \mathbf{n}_j^2, \mathbf{n}_k^2) \geq 0 \\ 1/|\mathcal{M}_{ijk}| & \text{otherwise} \end{cases} \tag{4.11}$$

where $\mathbf{n}_i \in \mathbb{R}^3$ denotes the normal at point i, and $\det(\mathbf{n}_i, \mathbf{n}_j, \mathbf{n}_k)$ denotes the determinant of the 3×3 matrix $[\mathbf{n}_i, \mathbf{n}_j, \mathbf{n}_k]$. This is introduced as a soft constraint in our framework, because in the extreme case, the normal could reverse its orientations when the surface undergoes very large deformations.

Finally, the third-order potential for each possible triple matching $(p_i^1, p_j^1, p_k^1) \rightarrow (p_i^2, p_j^2, p_k^2)$ is defined as a a weighted sum of the two types of potentials, *i.e.*,

$$\theta_{ijk} = \lambda_1 \theta_{ijk}^{\text{Möbius}} + \lambda_2 \theta_{ijk}^{\text{Gaussian}} \tag{4.12}$$

4.2.3 Dual-decomposition-based Optimization

As stated in the introduction (section 4.1), we aim to adopt the same insight as that of [Torresani *et al.* 2008] and recent order-reduction techniques [Ishikawa 2009] to deal with high-order graph-matching problems. We have reviewed the *dual-decomposition* MRF optimization framework [Bertsekas 1999, Komodakis *et al.* 2007a] in section 2.2.3, whose key idea is to decompose the original problem as a set of several sub-problems that are easier to solve. For the graph matching problem in Eq. 4.4, let $\boldsymbol{\theta}$ denote the vector of all the singleton, pairwise and triplet potentials, and $\mathcal{S}$ denote the set of subproblems. The decomposition of the original problem with objective function $E(\mathbf{x}; \boldsymbol{\theta})$ can be represented by:

$$E(\mathbf{x}; \boldsymbol{\theta}) = \sum_{s \in \mathcal{S}} \rho_\sigma E^s(\mathbf{x}|\boldsymbol{\theta}^s) \tag{4.13}$$

where ρ_s denotes the weight for each subproblem, $E^s(\mathbf{x}|\boldsymbol{\theta}^s)$ denotes the objective function of each subproblem s and the potential vectors $(\boldsymbol{\theta}^s)_{s \in \mathcal{S}}$ of the subproblems satisfy the following decomposition constraint:

$$\sum_{s \in \mathcal{S}} \rho_\sigma \boldsymbol{\theta}^s = \boldsymbol{\theta} \tag{4.14}$$

The lower bound $\Phi_s(\boldsymbol{\theta}^s)$ of each subproblem, *i.e.*, $\Phi_s(\boldsymbol{\theta}^s) \leq \min_{\mathbf{x}} E^s(\mathbf{x}|\boldsymbol{\theta}^s)$, constitute a lower bound for the original problem, *i.e.*,

$$\Phi(\boldsymbol{\theta}) = \sum_{s \in \mathcal{S}} \rho_s \Phi_s(\boldsymbol{\theta}^s) \leq \sum_{s \in \mathcal{S}} \rho_s E^s(\mathbf{x}|\boldsymbol{\theta}^s) = E(\mathbf{x}; \boldsymbol{\theta}) \tag{4.15}$$

In particular, we decompose the original problem into the following three subproblems:

1. a **linear subproblem** which considers only the singleton term $\sum_{a \in \mathcal{A}} \theta_a x_a$. This is a *linear assignment problem* and can be solved in polynomial time using for example the Hungarian algorithm (see section 4.1).

2. a **higher-order pseudo-boolean subproblem** where the high-order terms in Eq. 4.4 are reduced to quadratic terms [Boros & Hammer 2002] which can be solved by QPBO techniques [Kolmogorov & Rother 2007]. Regarding the order reduction, we employ the efficient method proposed in [Ishikawa 2009].

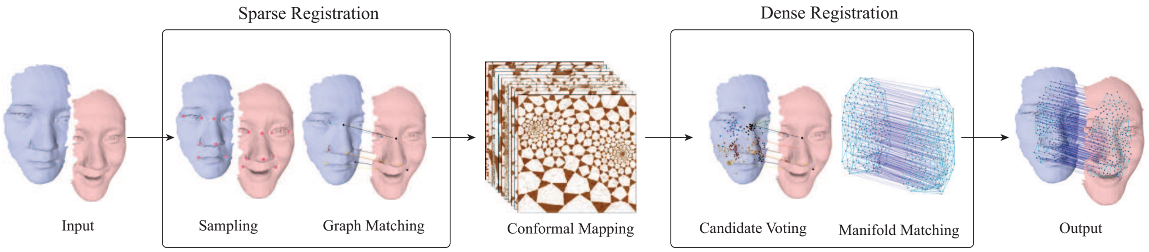


Figure 4.3: The Outline of the Algorithmic Framework for Surface Matching

3. **local subproblems** which divide the original surface into small regions and uses an exhaustive search to find the optimal matching solution in each small surface region.

The linear subproblem and the local subproblems used in the experiments are similar to those of [Torresani *et al.* 2008]. Besides, a higher-order pseudo-boolean subproblem is introduced to deal with the higher-order terms in Eq. 4.4. After solving the subproblems, the dual variables $\{\boldsymbol{\theta}^s\}$ are updated using a projected subgradient method as described in [Torresani *et al.* 2008] to maximize the lower bound $\Phi(\boldsymbol{\theta})$.

4.3 Towards Dense Surface Matching

The number of vertices n considered in this high-order formulation is the main computational bottleneck of our approach. In particular, when n becomes large, as in the case of dense surface matching, it is computationally expensive to solve the high-order graph matching problem presented above. Furthermore, the accuracy of the obtained solution degrade since the assumption of isometry is only an approximation and the distortion measurement based on Möbius energy becomes less discriminating when feature points are very close to each other. The graph structure of the above matching problem would also be very complex if we consider all possible triplets. Several heuristic ways were considered to prune off some triplets, such as restricting the number of triangles per vertex [Duchenne *et al.* 2009]. However, because of the complexity of the problem, such pruning schemes often lead to erroneous matching results when the number of feature points is large. Hence, towards dense surface matching, we propose a two-stage optimization pipeline which consists of *sparse feature matching* and *dense point matching*, as illustrated in Fig. 4.3.

In the sparse feature matching stage, an initial set of sparse feature points are selected among the local maxima of Gaussian curvature [Lipman & Funkhouser 2009] on the input surfaces S_1 and S_2. Using our high-order graph matching algorithm in section 4.2, we can compute the n_s correspondences between the two feature sets $\{p_1^1, p_2^1 \ldots, p_{n_s}^1\} \rightarrow$

$\{p^2_1, p^2_2, \ldots p^2_{n_s}\}$, where p^1_i and p^2_i $(i = 1 \ldots n_s)$ denote a pair of matched feature points on S_1 and S_2, respectively. In this stage, we only select a small set of feature points (typically $8 \sim 15$ in our experiments), so that the computational cost is low on finding the sparse correspondences and computing the associated conformal maps.

Since the initial feature points are selected among the vertices and the middle points of the edges of the meshes, the matching results could be unreliable if the mesh resolution is low. To address the above issue, we consider all conformal maps induced by different Möbius transformations, which are determined by every three correspondences between two surfaces, for the dense point matching.

4.3.1 Candidate Selection and Clustering

Candidate Selection

From the sparse matching stage, we have a set of sparse correspondences $\{p^1_1, p^1_2 \ldots, p^1_{n_s}\} \rightarrow \{p^2_1, p^2_2, \ldots p^2_{n_s}\}$ between S_1 and S_2. Because the surface deformation might not be isometric, we propose a candidate selection scheme based on Möbius transformations to compensate for the approximation error. Given any three correspondence pairs, $\{p^1_i, p^1_j, p^1_k\} \rightarrow \{p^2_i, p^2_j, p^2_k\}$, the corresponding Möbius transformation can be computed very efficiently in a closed form [Lipman & Funkhouser 2009]. Under such a Möbius transformation, any point $p^1 \in S_1$ will be mapped to a different candidate location $c(p^1) \in S_2$. Thus, for each point on the source surface, we can compute its candidate locations in the target surface by considering all possible Möbius transformations from the feature correspondences. Please note that our candidate selection approach differs from the Möbius voting method [Lipman & Funkhouser 2009] in two ways: (1) our method computes the positions of matching candidates for each dense point rather than finding sparse feature correspondences; and (2) multiple clusters are computed from the candidate positions of each point and used to obtain a dense matching result.

One advantage of our candidate selection approach is robustness. If any part of the sparse matching result is accurate, the matching candidates given by the Möbius groups will distribute closely around the true location for surfaces undergoing near-isometric deformations. Another advantage is that this scheme provides a fast and effective way of constraining the search space for any point on the surface.

Candidate clustering

Based on the above candidate locations, we want to use the underlying distribution to reduce our search space for the dense matching. It is also important that the dense matching should optimize the same objective as in the sparse matching stage. For any matching candidate point $c(p^1) \in S_2$ of a source point $p^1 \in S_1$ that is obtained by aligning three correspondences $\{p_i^1, p_j^1, p_k^1\} \rightarrow \{p_i^2, p_j^2, p_k^2\}(i, j, k = 1 \dots n)$, there is a cost $\theta_{ijk}^{\text{Möbius}}$ in the matching energy of Eq. 4.4. Intuitively, the lower the value of $\theta_{ijk}^{\text{Möbius}}$ and the closer the curvature and texture is, the more likely p^1 and $c(p^1)$ match. Therefore, we define the likelihood of each candidate matching $p^1 \rightarrow c(p^1)$ under the alignment of $\{p_i^1, p_j^1, p_k^1\} \rightarrow \{p_i^2, p_j^2, p_k^2\}$ as follows

$$f_{ijk}(p^1, c(p^1)) = e^{-\theta_{ijk}^{\text{Möbius}}} \tag{4.16}$$

where $\theta_{ijk}^{\text{Möbius}}$ is the Möbius matching potential in Eq. 4.10. To obtain the candidate distribution for each point $p^1 \in S_1$, we use a kernel density estimate (KDE) with the density function defined as

$$\rho(p^1, c(p^1)) = \sum_c f_{ijk}(p^1, c(p^1)) K(\frac{\|c(p^1) - c(p_c^1)\|}{h}) \tag{4.17}$$

where $c(p_c^1)$ is the center location of each kernel K in S_2 and h is the kernel bandwidth. The mean shift clustering [Comaniciu & Meer 2002] is employed to find the modes of this density. Compared to parametric representations, KDE does not require nonlinear optimization to learn the distribution parameters.

Since we search for the modes in Eq. 4.17 on the 2D manifold instead of the 3D Euclidean embedding space, the distance function should be defined as the geodesic distance on the surface. However, as illustrated in Fig. 4.4 most of the candidate locations are close to the center, so the Euclidean distance was used in our experiments to simplify the mode search. To handle partial surface matching, we only select the modes with density higher than 0.1 and the closest point on the surface as the final matching candidates. If no such mode exists, we report that there is no reliable correspondence point. The average number of resulting matching candidates in our experiments is $1 \sim 6$. So our candidate selection and clustering method can significantly reduce the search space.

4.3.2 Local High-order Graph Matching

Based on the matching candidates obtained for each vertex, our goal now is to find a good matching position locally for each dense point. This problem can be formulated similarly to the high-order graph matching problem defined in section 4.2.1. Since the candidate selection scheme in section 4.3.1 has removed the ambiguities caused by the Möbius transformations, we only need to consider the matching cost based on texture and geometric similarities defined in Eq. 4.7, as well as the orientation consistency imposing that each triangle $\triangle_{p_1 p_2 p_3}$ and its matched triangle $\triangle_{p'_1 p'_2 p'_3}$ should have the same orientation in the uniformization domain, which is known as having no flip in [Sheffer *et al.* 2006].

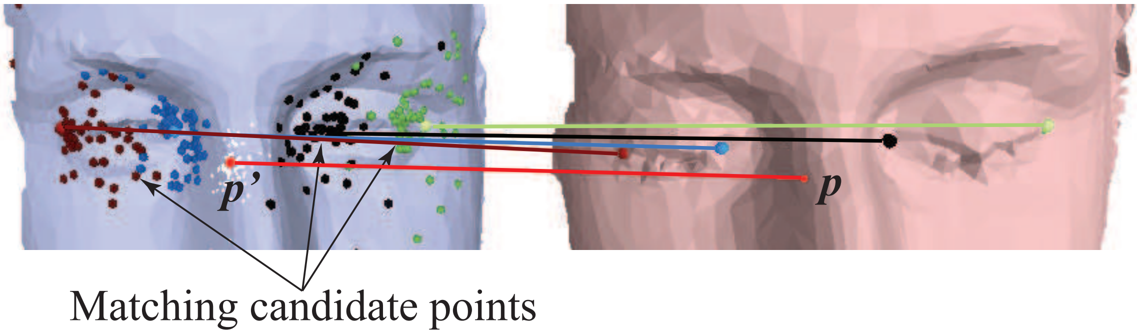


Figure 4.4: Example of Candidate Selection and Clustering. The figure shows the matching candidate points from different Möbius transformations and clustering. For any point p from the source surface, the clustering of candidates on the target surface gives us final matching candidates.

More specifically, for the three vertices of each triangle $\triangle_{123}$, we define the potential of matching (p_1, p'_1), (p_2, p'_2) and (p_3, p'_3) as follows

$$\theta_{123,1'2'3'} = \begin{cases} \theta^{\infty} & \text{sign}(\triangle_{123}) \neq \text{sign}(\triangle_{1'2'3'}) \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

where θ^{∞} is a sufficiently large number. sign$(\triangle_{123})$ and sign$(\triangle_{1'2'3'})$ denote the orientation of the triangle $p_1p_2p_3$ and $p'_1p'_2p'_3$, respectively, in the uniformization domain. From the candidate clustering, it is not guaranteed that every point has at least one matching candidate. Therefore, we remove the points without any matching candidate and obtain a triangulation for the remaining points on S_1 through the Delaunay triangulation algorithm [de Berg *et al.* 2000] in the uniformization domain.

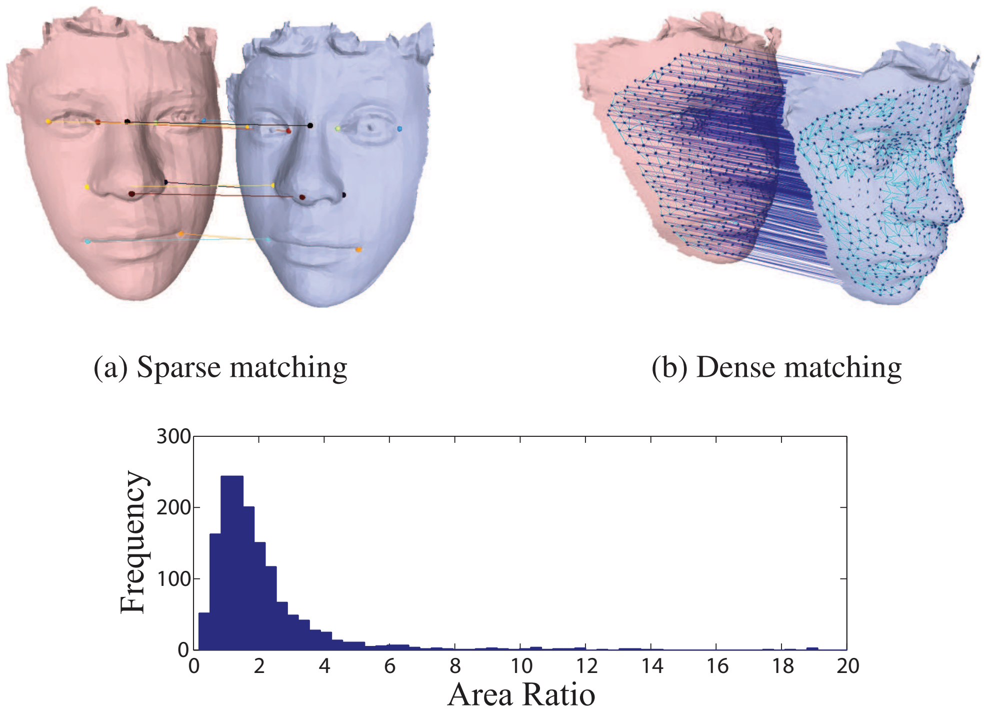
Suppose for each point $p \in S_1$, its matching candidates are given by $\mathcal{C}_p = \{p_i | p_i \in S_2, i = 1, 2, \ldots, n_p\}$. We define the boolean indicator variable:

$$x_p^i = \begin{cases} 1 & \text{if } p, p_i \in \mathcal{C}_p \text{ are active correspondences} \\ 0 & \text{otherwise.} \end{cases} \tag{4.19}$$

Assuming that each $p \in S_1$ is matched to at most one of its candidates, we have the matching constraint:

$$\sum_{p_i \in \mathcal{C}_p} x_p^i \leq 1 \tag{4.20}$$

Therefore, the same optimization technique as described in section 4.2.3 can be applied to solve the above problem.



(a) Sparse matching

(b) Dense matching

(c) Matching area ratio histogram

Figure 4.5: Example Face Matching Result: (matched/total = $2098/2644$)

Compared to the graph matching problem in section 4.2.3, one major advantage of the local graph matching algorithm is that the number of matching candidates for each point is typically less than 6 and, therefore, the number of variables is very small. In particular, to match n points locally, there are only $O(n)$ variables and $O(n)$ triplet terms since the dense points are triangulated in the planar parametric domain.

4.4 Experimental Results

4.4.1 Experimental Setting

Our algorithm is implemented on an Intel® Xeon(TM) 3.4G PC with 4G RAM and an NVIDIA® Geforce 9800GTX$+$ graphics card. We developed a matching plugin for the open source software Meshlab[2]. For the mean shift algorithm, we used the source code available online[3]. For the potential functions of the graph matching algorithm defined in section 4.2.2, the weights of Eq. 4.7 and 4.12 are defined as $\lambda_0 = 1$, $\lambda_1 = 0.1$ and

[2] http://meshlab.sourceforge.net/

[3] http://www.caip.rutgers.edu/riul/research/code.html

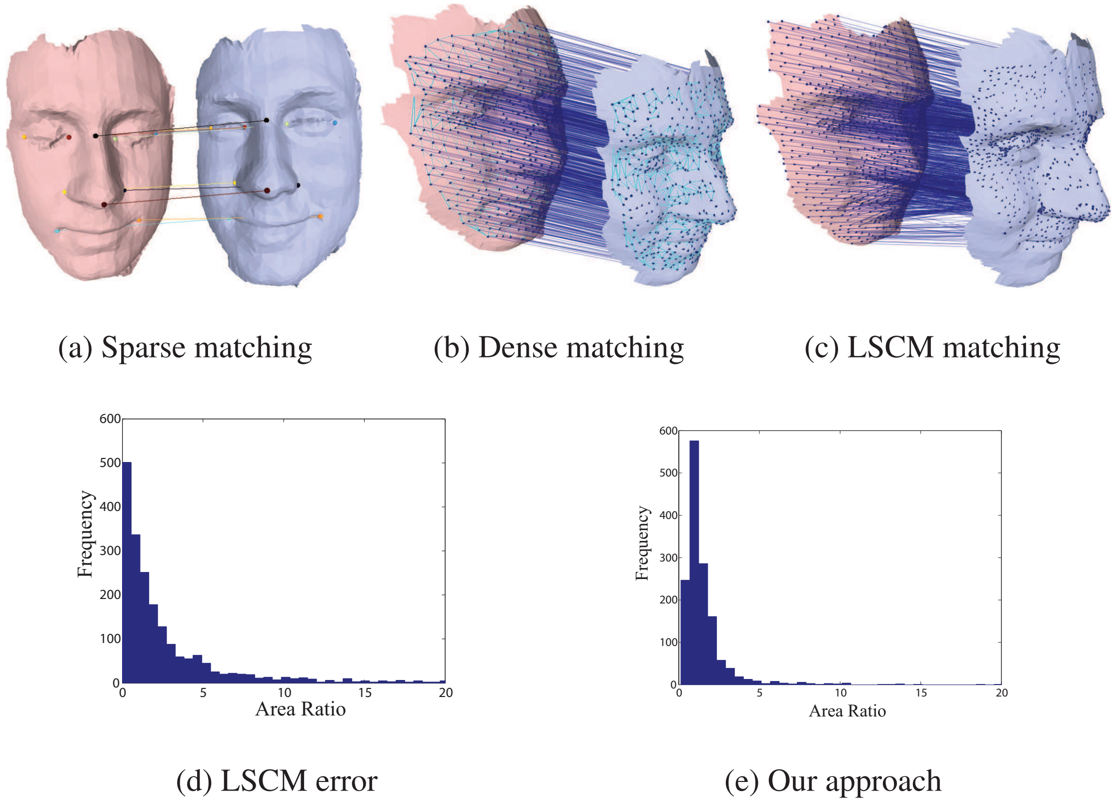
$\lambda_2 = 1$, and the kernel bandwidth of Eq. 4.17 is set to be 0.01 times the diameter of the target surface. The mid-edge uniformization algorithm was used for the conformal mapping [Lipman & Funkhouser 2009, Pinkall & Polthier 1993]. The computation of mid-edge uniformization involves solving a symmetric linear equation, which can be efficiently computed by GPU [Buatois *et al.* 2009]. For a mesh with 10^4 faces the computation takes less than 1 second.

Since we consider almost all the triplets, the graph complexity scales cubically without pruning. Therefore, rather than searching for more sparse feature correspondences in the first stage, we try to find more accurate matching results for a few features. For example, 10 sparse feature correspondences will give us 120 matching candidate positions for each point which are enough for finding final candidate points. To match 10 feature points, the graph encoding step takes around 5 minutes and the graph matching step takes less than 1 minute. The candidate selection and local high-order graph matching of 10^3 points based on the 10 sparse features takes around 1 minutes. Compared to previous work [Lipman & Funkhouser 2009, Tevs *et al.* 2009] which only computes around 100 correspondences, our algorithm not only runs faster but also achieves more correspondences. For the high-order graph matching algorithm in section 4.2.3, the convergence of the dual-decomposition optimization depends on the input features. In our experiments, we observed that the more outliers (un-matched points), the more iterations it took to converge.

4.4.2 Results

We evaluate our new algorithmic framework using a number of challenging data. In our experiments, we match surfaces with large deformations and inconsistent boundaries (partial overlapping). The number of vertices for each mesh is in the range of $1,500 \sim 4,000$. With our high-order graph matching algorithm, we can find the dense matching for $60 \sim 90$ percent of all vertices, which is illustrated as matched/total (no. of matched vertices/no. of total vertices of the source surface) for each example. The lion data of Fig. 4.1 comes from [Sumner & Popović 2004] and the face and hand data are captured with texture by the 3D scanner introduced in [Wang *et al.* 2005]. To measure the quality of dense registration, from the Delaunay triangulation of the points on the source surface, we consider the ratio of the area of each local triangle to the area of its matched triangle. For the natural deformations (*e.g.*, expression change, stretched arms or bending figures) we experimented with, the local area is not expected to undergo abrupt change. Therefore the area ratio is expected to be close to one for every local triangle.

Matching with largely inconsistent boundaries and partial overlapping: The mid-edge uniformization algorithm allows to map the boundaries of the surface to slits and preserve


(a) Sparse matching

(b) Dense matching

(c) LSCM matching

(d) LSCM error

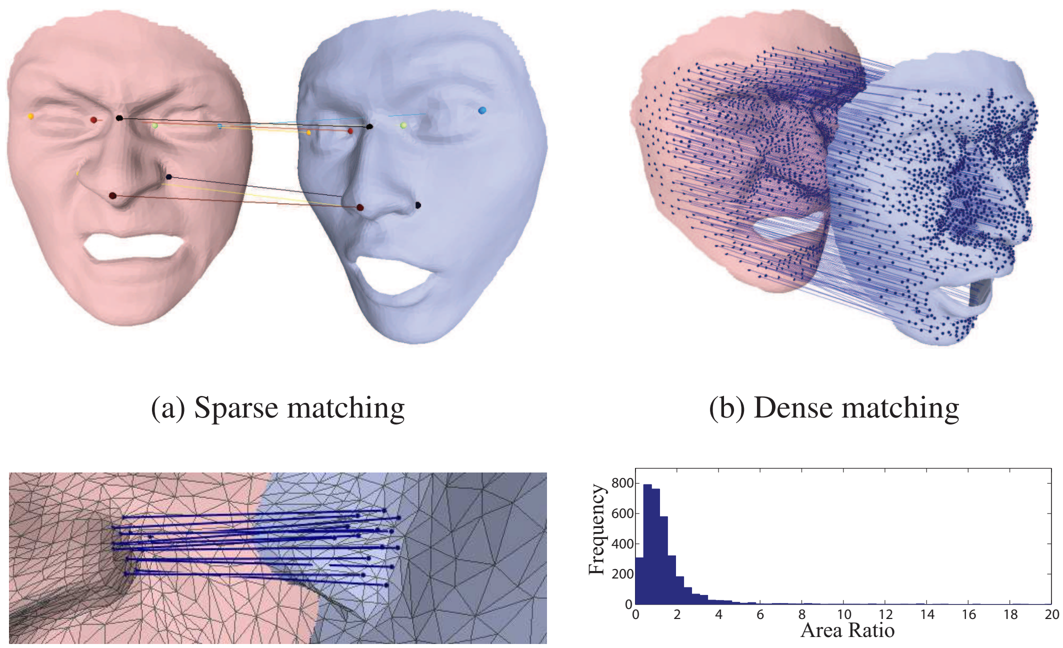
(e) Our approach

Figure 4.6: Comparison with LSCM Approach [Wang *et al.* 2007]. (matched/total = $1455/1635$). Notice the high number of flipped triangles in (c)

the conformal structure of the surface in an exact sense. Hence it is suitable for matching partially overlapping surfaces. This property can be combined with our candidate selection scheme to determine the outliers near the boundary where the mean shift clustering returns a low score. Examples are shown in Fig. 4.5, 4.6, and 4.7. An example of significant non-overlap between the two meshes is shown in Fig. 4.3.

Matching with large deformations: Fig. 4.7 and 4.8 show results that match two surfaces undergoing a large deformation. Even when the sparse features can not all be selected consistently (as shown in Fig. 4.8), our high-order graph matching algorithm in section 4.2.3 is able to find reliable sparse correspondences (Fig. 4.8(a)) and obtain a dense surface matching result (Fig. 4.8(b)) through the two-stage optimization scheme described in section 4.3.

Comparison experiments: Fig. 4.6 shows a comparison between our algorithm and the least square conformal mapping (LSCM) approach [Wang *et al.* 2007]. Although LSCM can handle free boundaries, there is no theoretical guarantee that the conformal structure is preserved near the boundary and it might include self-intersections in the mapping [Sheffer *et al.* 2006]. In our comparison, we use the feature correspondences com-



(a) Sparse matching

(b) Dense matching

(c) Closeup of the dense matching

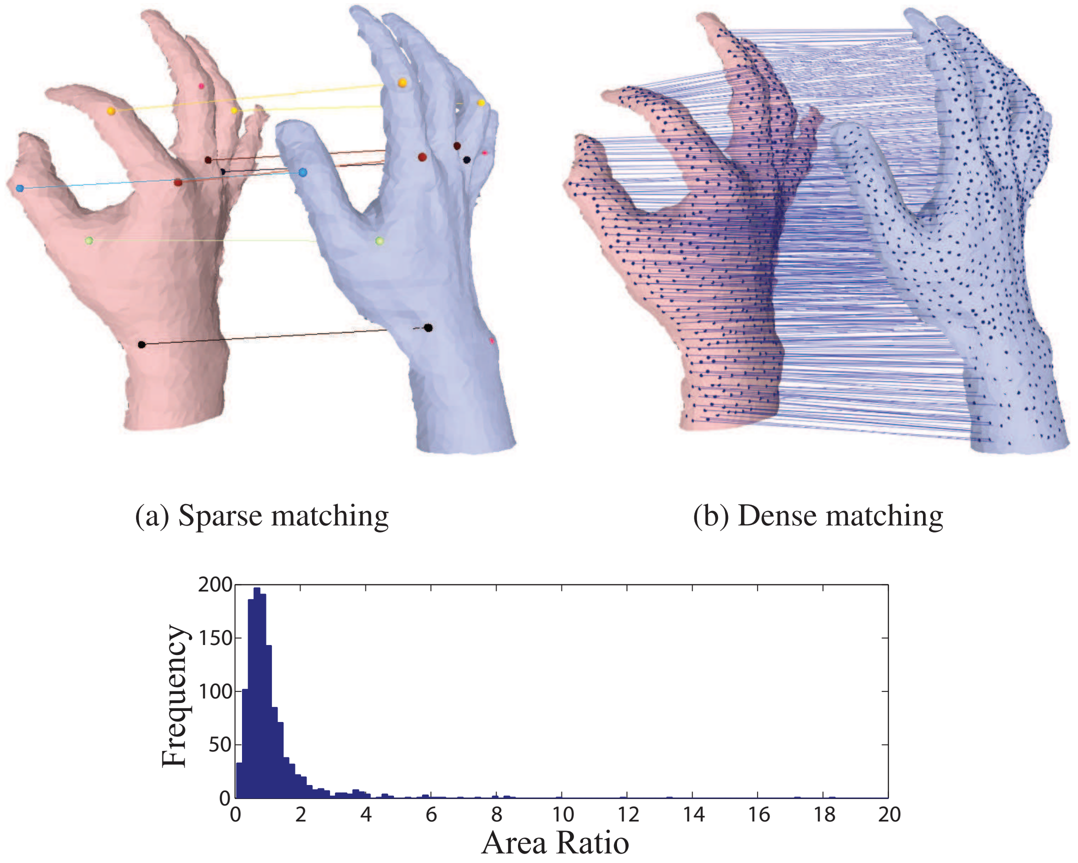
(d) Matching area ration histogram

Figure 4.7: Dense Matching under Large Non-rigid Deformations. (matched/total = $2378/3633$)

puted from the sparse matching stage to initialize the LSCM experiments. The inaccuracy of the LSCM approach can be observed in Fig. 4.6(c). In this example, although all vertices on the left mesh are matched to the right mesh, there are approximately 42 percent flipped triangles. Note that here we cannot compare directly with the results in [Wang *et al.* 2007] where the initial feature points were manually selected.

4.5 Conclusion

In this chapter, we have proposed an algorithmic framework for non-rigid surface matching. In particular, a high-order graph matching formulation is used to combine local distortion regarding the appearance and geometry similarity as well as global structure distortion (*i.e.*, intrinsic deformation errors) between deformed surfaces, resulting in a robust algorithm to establish sparse matching between two non-rigid surfaces with large deformations, partial matching and inconsistent boundaries and scales. Furthermore, towards achieving dense surface matching, a two-stage scheme has also been introduced to constrain the search space through candidate selection and local graph matching. The whole method is modular with respect to the potentials used to determine optimal partial corre-



(a) Sparse matching

(b) Dense matching

(c) Matching area ratio histogram

Figure 4.8: Dense Matching under Multiple Articulated Deformations. (matched/total = 1224/1786)

spondences.

While isometry is a good approximation to many real-world deformations, there are also many other types of deformations that do not fall into this category. An important case is the variability within a class of shapes (*e.g.*, fat or thin man). The modeling of such a variability is extremely important for many computer vision and medical imaging problems where a common model is used to represent the instances of an object class, such as knowledge-based image segmentation and 3D model reconstruction from 2D views. In such cases, one usually resorts to statistical modeling to deal with such intra-class shape variations. To this end, we have studied the statistical shape modeling and applications based on it, which will be presented in the next chapter (chapter 5).

Chapter 5

3D Model Inference from 3D/2D Images

In this chapter, we aim at developing graph-based models for 3D model inference without explicit estimation of global parameters (*i.e.*, the global pose of the object of interest or the camera viewpoint). To this end, we first propose a pose-invariant shape prior model that can be naturally encoded within higher-order clique potentials. Based on this shape model, we introduce a single-shot optimization framework for knowledge-based image segmentation of challenging medical image data using a higher-order MRF, where a dual-decomposition-based method is used to recover the optimal solution. This approach has been validated through challenging experiments on segmentation of human skeletal muscles. Furthermore, in order to partially address the influence of camera pose in visual perception, we propose a unified higher-order MRF formulation to simultaneously determine both the optimal 3D landmark model and the corresponding 2D projections without explicit estimation of the camera viewpoint, which is also able to deal with misdetections as well as partial occlusions. Promising results on standard face benchmarks demonstrate the potential of this approach.

5.1 Introduction

Low level segmentation and primitive-based tracking as studied in chapter 3 serve as core components to solutions of many computer vision problems. Despite their strength, their applicability is limited though to low or mid-level vision since in general either a more precise delineation of the object of interest or estimation of dense motion fields is required. In such a context, simplistic priors as the one employed in the previous chapter fail short with respect to the expected performance. Introducing such priors can happen either in the 2D space or directly on the 3D world.

Segmentation with shape priors often requires a learning stage where given a set of training examples one seeks for a probabilistic representation of the observed variation. To this end, all training examples are first brought to the same reference space (*e.g.*, through linear registration) and then relative deformations with respect to the average shape are modeled. Given such a prior model, segmentation aims at recovering the best possible instance of the learned manifold in the image space. Such a process requires bringing the observed image to the same reference space used during learning. This is usually achieved through a linear extraction/registration of the mean shape to the observed image. Then, combination of prior knowledge and image support are used to delineated the optimal shape. Such an approach has been extensively used in computer vision but suffers from the need of registering all examples to a reference space, which introduces a strong bias and results on a sequential optimization method that can be very sensitive approach.

During the past two decades, significant effort has been carried out towards appropriate modeling of shape variations in the 2D space. Such an approach can mostly cope with known viewpoint object configurations and aims at modeling variability of a population of exemplars which have been mostly captured from the same viewpoint. However, coping with severe viewpoint differences often requires modeling the variations of the shape of interest in the original 3D space. This eliminates the viewpoint issue with regards to the 2D alternative and could lead to better expression of the shape manifold. On the other hand, it introduces during inference (especially when considering 2D images that is often the case), the need of estimating the projection matrix between the 3D model and the corresponding image. In the most general case, such a configuration is unknown and the advantage of modeling directly the 3D variation is compromised from the need of estimating the camera parameters. The problem is often solved sequentially or in an alternating manner, first the projection parameters are estimation, then segmentation is solved that is fed back to the viewpoint estimation process.

Numerous efforts have been carried out towards proper modeling of shape variations. In both problems above, a well-established limitation of coordinate-descent approaches is that they provide no guarantee on the optimality of the estimation and are prone to be trapped in local minima. In the rest of this introduction, we give a detailed description of the context and motivations of the approaches that we will develop here.

5.1.1 Knowledge-based Segmentation

Image segmentation is a fundamental problem in computer vision and medical image analysis. Such a problem is intrinsically ill-posed and the use of prior knowledge is often considered to address it. In particular, the integration of prior knowledge is very important when extracting specific objects from observed images, towards achieving superior performance and high robustness to challenging cases where noise, occlusions and low-contrast are present in the images.

Knowledge-based segmentation consists in recovering a region of interest in an observed image and generally involves three main parts: *shape representation*, *prior learning* and *inference*. First of all, one has to choose an appropriate representation for modeling the shape of the object of interest. Once the shape representation is determined, training examples are used to learn statistics on the shape model which is referred to as a *statistical shape model (SSD)*. Then in the segmentation stage, the inference of the shape model is done by seeking a compromise between data-attraction and the fitness to the prior model.

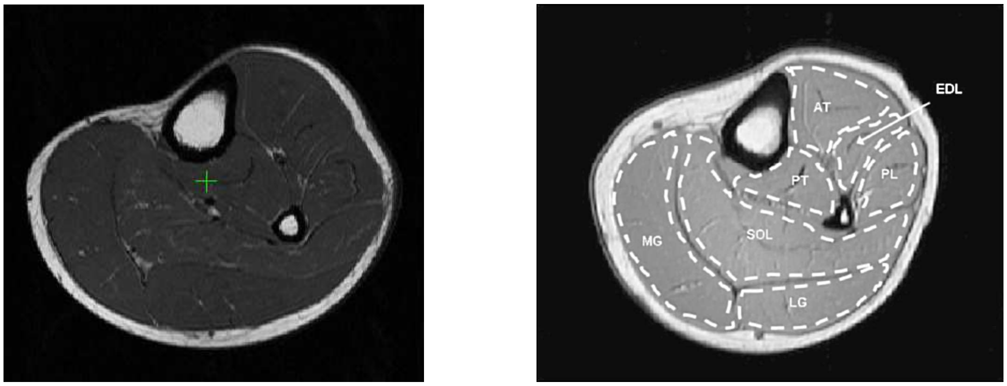
Statistical Shape Models

There are diverse representations for modeling the shape of an object, such as landmark-based models (also referred to as *point distribution models (PDMs)*) [Cootes *et al.* 1995, Cootes *et al.* 2001], *level set* representations (often referred to as *implicit representations*) [Osher & Fedkiw 2002, Cremers *et al.* 2007], medial models [Blum 1973, Pizer *et al.* 2003], frequency-domain representations [Staib & Duncan 1996, Essafi *et al.* 2009b] and articulated models [Sigal & Black 2006a, de La Gorce *et al.* 2011]. In point distribution models, the shape is represented using a set of control points (often corresponding to landmarks) distributed on the surface. The coordinates of all the points are concatenated into a vector $\mathbf{x}$ so that the value of $\mathbf{x}$ determines the shape. In implicit representations, the boundary of the shape is embedded in a high dimensional space (*e.g.*, *signed distance map* [Osher & Fedkiw 2002]) and is characterized by the zero level set. Medial models characterize a shape using its medial axis and the corresponding radii of the bi-tangent spheres. Frequency-domain representations refer to a set of techniques which apply Fourier transform or wavelet transform on the shape and describe the shape in the frequency-domain. Articulated models are employed to represent objects such as the human body and the hand, by capturing the kinematic constraints between neighbor components.

The objects of interest in the works presented in this chapter are non-articulated objects such as muscles and the human face. In such a context, we are specially interested in PDMs, since the landmarks involved in a PDM can be naturally modeled as nodes in graphical models, while other representations are difficult to be modeled using graphs. The most well-known PDMs are *active shape models (ASMs)* and *active appearance models (AAMs)*, which were proposed in [Cootes *et al.* 1995] and [Cootes *et al.* 2001], respectively. Such models are constructed in two steps: i) during the first stage, all the training samples are aligned in a common coordinate frame using for example *Procrustes Analysis* [Dryden & Mardia 1998]; ii) then, a dimensionality reduction is performed using *Principal component analysis (PCA)* [Jolliffe 2002] so as to obtain a limited number of modes that can best capture the most important variations present in the training data. They offer a good compromise between computational complexity and model expressiveness potential and have therefore been widely used in the literature.

However, knowledge-based segmentation methods that use such global statistical models and many others often exhibit two important limitations. The first limitation lies in the fact that the shape prior cannot be pose-invariant since it is learned in a certain coordinate frame, as mentioned earlier in the introduction. Thus, the estimation of the global pose (translation, rotation and scale) is required both in the training and in the inference stages. Such methods may introduce a certain bias on the segmentation process since data are often to be registered in the reference space. More importantly, since the estimation of the global pose is usually done by a local search, these methods are prone to fail if the initialization is far from the ground-truth pose. The second limitation is related to their ability to capture statistics and variations on high-dimensional spaces from a small number of training examples, due to the global representation of the shape as well as the linearity of the models. The samples-vs-dimensionality ratio of representations is also a well-known problem in medical imaging, due to the fact that the number of available training data with ground truth shapes is often very limited.

Various segmentation methods have been proposed aiming to partially deal with such limitations. For example, non-linear statistical models have been investigated to in order to better capture shape variations. One can cite for example the PDMs based on mixture models (*e.g.*, [Cootes & Taylor 1999, Gu & Kanade 2008]), kernel PCA [Scholkopf *et al.* 1998] (*e.g.*, [Romdhani *et al.* 1999, Twining & Taylor 2001]) and the *Gaussian process latent variable model (GPLVM)* [Lawrence 2004] (*e.g.*, [Chen *et al.* 2010, Huang *et al.* 2011]). At the same time, various works have been done to develop shape models based on local interactions between control points. [Seghers *et al.* 2007a] introduced a 2D shape model that is represented by a closed curve consisting of a sequence of landmarks. The prior is encoded by the statistics on three kinds of measures based on the Euclidean distances between two successive landmarks or the relative positions of three successive landmarks. Such statistics inherit different invariance properties such as translation-invariance and translation/rotation-invariance. Due to the chain structure of the shape model, dynamic programming [Bellman 1957, Cormen *et al.* 2009] was adopted as the inference algorithm. However, such an approach is not able to deal with 3D cases and only using constraints based on neighbor landmarks cannot capture well the underlying shape manifold. The translation-invariant prior of [Seghers *et al.* 2007a] was also employed in [Seghers *et al.* 2007b] to address the 3D segmentation of the liver from contrast enhanced CT images, through a heuristic search method. In this approach, observed images have to be registered to the reference image of the training set before the segmentation processing due to the fact that the used prior is not pose-invariant. Recently, a PDM was proposed in [Besbes *et al.* 2009] towards knowledge-based segmentation, where the prior information



(a) A slice of muscle data (b) Manual expert segmentation of muscles

Figure 5.1: MRI Data of Calf Muscles (courtesy [Essafi *et al.* 2009a]).

about the shape is expressed through a combination of local interactions. More specifically, the Euclidean distance between pairs of landmarks are normalized by the scale of the objects (*i.e.*, the sum of distances between all the pairs of landmarks) and then statistics are built on such normalized distances. Such a prior can be naturally encoded using the pairwise potentials of an MRF. On the other hand, the data likelihood is decomposed (via *Voronoi diagram* [Aurenhammer 1991]) into a sum of local terms that are encoded in the singleton potentials of the MRF. In this way, the segmentation problem is formulated as a MAP inference in the MRF model. This method has shown to outperform standard methods such as AAMs. As a global representation (where we know the position of all the points), such a prior model is pose-invariant (translation, rotation and scale). However, it is still not "intrinsically" scale-invariant and cannot be exactly factorized into an MRF, since the definition of every local term depends on the scale of the object, which requires the estimation of the sum of the distances between all the pairs of points and thus depends on the positions of all the points of the shape model. To deal with this, an iterative scheme was employed in [Besbes *et al.* 2009], where the shape model is deformed gradually during the evolution and at each iteration, the scale is estimated using the configuration of the shape model at previous iteration. The performance of such a method depends on the quality of the scale approximation obtained during the iterative search. In this work, we aim to search for a statistical shape model that is intrinsically pose-invariant.

Another motivation for developing a pose-invariant prior came from applications in medical imaging. Medical imaging provides a variety of image acquisition techniques such as radiography, tomography, ultrasound, magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI), to visualize the human body for clinical and medical research purposes. Segmentation is certainly one of the most important medical image processes required for clinical examination and biological analysis. However, in comparison to natural images, medical modalities often yield certain special difficulties in the segmentation task.

One major difficulty is related to the choice of data terms. Data terms used in image segmentation are generally based on edges (*e.g.*, [Kass *et al.* 1988, Brejl & Sonka 2000, Iannizzotto & Vita 2000]) and/or region driven (*e.g.*, [Rother *et al.* 2004, Kohli *et al.* 2008b, Boykov & Funka-Lea 2006, Paragios & Deriche 2002]). In the first case, one seeks to position the solution onto pixels exhibiting important intensity discontinuities, which is achieved through a weighted surface integral. Region-based methods assume that the object and the background have distinct statistical properties and seek to create a partition that maximizes the posterior probability density with respect to them. However, both strategies cannot handle satisfactorily anatomical cases where separate regions of interest can belong to the same class of tissue. In such cases conventional image support is lacking: edges are poorly informative and a statistical discrimination of regions would be bound to fail. Calf muscle MRI segmentation is a typical example (see Fig. 5.1), since there is no prominent difference of tissue properties between neighbor muscles and since tissue boundaries separate adjacent muscles only sparsely and heterogeneously. Therefore, medical segmentation issues, such as this of the calf muscles that was hardly studied in the literature [Blemker *et al.* 2007, Essafi *et al.* 2009a], provide a perfect frame to illustrate the benefit of an alternative image support. A natural way of building an adapted image support relies on landmark classifications, where feature vectors exploit the information around a particular location and exhibit highly discriminative capacity. This strategy has been previously employed in various medical image segmentation applications (*e.g.*, [Donner *et al.* 2007, Seghers *et al.* 2007a, Seghers *et al.* 2007b]), where a set of candidates are detected for each landmark and then prior information are fused in order to select the optimal candidate for every landmark. These facts have inspired and motivated us to develop a one-shot knowledge-based segmentation approach using a landmark-based image support, that is particularly adapted to the segmentation of the challenging medical cases such as MRI data of calf muscles. In order to be one-shot, such an approach would necessitate the pose-invariance property of the shape model.

Our Approach for Segmentation from 3D Images

We propose a novel segmentation approach that is able to address 3D segmentation, while being pose invariant and able to capture local variations from small training sets. The representation of the shape is a PDM which consists of a set of landmarks located on the boundary surface and determine the entire surface through conventional interpolation algorithms such as thin plate spline (TPS) [Bookstein 1989]. Prior knowledge is modeled through higher-order statistics on the PDM, which are invariant to similarity transform (*i.e.*, translation, rotation and scale) and can be learned from a small number of training examples. The entire manifold is described through the accumulation of such local constraints. Data likelihood is determined using the randomized forest [Breiman 2001] learning approach that provides an efficient classification algorithm for points of interest exhibiting certain statistical properties. Finally, the segmentation is formulated as a MAP inference in a higher-order MRF, where the pose-invariant priors are encoded within higher-order clique potentials and the data support is encoded in the singleton terms. Such an approach provides a one-shot optimization result that does not depend on initial conditions nor on the reference pose. In order to optimize the energy of the higher-order MRF, we adopt the dual-decomposition optimization framework [Bertsekas 1999, Komodakis *et al.* 2007a] (see section 2.2.3) and propose to decompose the original problem into a series of sub-problems each of which corresponds to a factor tree [Frey 1998, Bishop 2006] (see section 2.1.4). The inference in a factor tree can be done exactly in polynomial time using max-product belief propagation algorithm (see section 2.2.2). The performance of the method is evaluated in the challenging application of segmentation of the calf muscle, which demonstrates the potential of the proposed method.

5.1.2 3D Model Inference from Monocular 2D images

In a second stage, we consider 3D model inference from monocular 2D images, which is one of the most challenging problems in computer vision. This is due to the fact that both camera estimation and 3D model optimization have to be addressed within a single framework. In the most general case, the camera parameters are unknown, the 3D model itself usually inherits high complexity (high degrees of freedom even for non-articulated objects), while at the same time image features can be ambiguous because of noise and occlusions for instance. There are numerous applications involving the above scenario, such as articulated object pose estimation (*e.g.*, [O'Rourke & Badler 1980, Sigal *et al.* 2007, Forsyth *et al.* 2005, de La Gorce *et al.* 2011]), shape/surface estimation (*e.g.*, [Balan *et al.* 2007, Guan *et al.* 2009, Chen *et al.* 2010, Salzmann & Fua 2010]), facial analysis (*e.g.*, [Pighin *et al.* 1998, Blanz & Vetter 1999, Gu & Kanade 2006]), traffic monitoring with 3D model-based tracking (*e.g.*, [Roller *et al.* 1993, Mueller *et al.* 2003, Leotta & Mundy 2011], architecture modeling [Walker & Herman 1988, Simon *et al.* 2011] and medical imaging (*e.g.*, [Kurazume *et al.* 2009, Markelj *et al.* 2010]). Such an inference process usually involves three steps[1]: the first aims to determine a compact representation of the 3D model, the second to associate such a representation with the 2D observations, and the last to recover the optimal parameters of the model.

In section 5.1.1, we have reviewed briefly statistical shape models as well as diverse

[1] Note that there are also a kind of methods (*e.g.*, [Bregler *et al.* 2000, Sigal & Black 2006b]) which do first 2D estimation and then recover the 3D configurations.

representations for modeling non-rigid objects. We recall that the limitations of global statistical models that have been discussed in section 5.1.1 are still valid in the context of 3D model inference from monocular 2D images. Once a shape model has been built for a class of objects, the next steps consist of defining an image likelihood and combining it with the 3D model prior towards optimal estimation of the 3D model. Since the image likelihood is related to both the 3D model configuration and the camera parameters, the model estimation is often achieved through an alternating search, an EM-style approach or other local search methods (*e.g.*, [Sandhu *et al.* 2009, Gu & Kanade 2006]). Given an initial 3D-2D correspondence map, the camera parameters are first estimated and then used to define the fitting error between the model and the image. This error is to be optimized by gradient-driven methods and iterative search processes so as to estimate both the correspondences and the optimal model configuration. For example, a level set shape representation, together with the prior information learned using PCA, was used in [Sandhu *et al.* 2009] to deal with 3D model estimation and 2D image segmentation. The objective function is optimized by iteratively performing gradient descent with respect to the shape parameters (*i.e.*, PCA coefficients) and the pose parameters (corresponding to the camera viewpoint). A well-known PCA-based statistical model called *morphable model* was proposed in [Blanz & Vetter 1999] to model 3D shape and appearance and human face and to perform 3D reconstructions from 2D images, where the global pose is manually determined. [Gu & Kanade 2006] proposed an approach to deal with face alignment from a single 2D image. A 3D landmark-based face model is adopted to represent the face and a PCA-based prior on the 3D model is learned from synthetic training data. The deformation of such a model and the global 3D pose are adjusted iteratively via an EM-based approach to fit an observed image. Despite promising performance achieved by such approaches, the fact that an explicit estimation of the camera viewpoint parameters is required in the process is a major drawback, since coordinate-descent approaches are prone to be trapped in local minima and provide no guarantee on the optimality of the estimations, which would require a good initializations of the 3D model and/or the camera configuration before the optimization process.

Such a context led to the problem that is addressed in this work, which consists of the estimation of 3D models from 2D images without explicit estimation of the camera viewpoints. As a first milestone towards this goal, we aimed to develop a unified approach for viewpoint 3D landmark model inference from monocular 2D images based on the previously proposed segmentation formulation (see section 5.2).

Our Approach for 3D Landmark Inference from 2D Images

We propose a novel one-shot optimization approach to simultaneously determine both the optimal 3D landmark model and the corresponding 2D projections without explicit estimation of the camera viewpoint, which is also able to deal with misdetections as well as partial occlusions. To this end, we formulate the problem as a maximum a posteriori (MAP) estimation task which involves 3D pose parameters, associated 2D correspondences and visibility states. We derive a posterior probability as the product of an image likelihood, a visibility prior, a 3D geometric prior and a projection consistency prior constraining the 2D and 3D configurations. In order to circumvent the need of viewpoint estimation, we adopt a high-order decomposition of the 3D model that enables to determine the projection error between a given 3D configuration and the corresponding 2D landmark positions in a distributed manner. Furthermore, an explicit visibility modeling is also introduced to cope with misdetections and outliers. The MAP inference is then naturally transformed into a higher-order MRF optimization problem and all the latent variables are inferred through a dual-decomposition-based method. The proposed formulation has been validated in the context of 3D facial pose estimation from 2D images. Promising results on standard face benchmarks demonstrate the potential of our method.

5.1.3 Outline of the Chapter

The remainder of this chapter is organized as follows. First in section 5.2, we present the method that deals with knowledge-based 3D image segmentation. This presentation also includes our pose-invariant shape model involved in the formulations proposed for both problems. Section 5.3 is dedicated to the simultaneous estimation of a 3D landmark model and of 2D correspondences. The used higher-order MRF optimization approach and experimental validation of both methods are presented in section 5.4. Finally, we conclude this chapter in section 5.5.

5.2 Knowledge-based Segmentation Using Pose-invariant Priors

5.2.1 Pose-invariant Shape Modeling

The shape model consists of a set $\mathcal{V}$ of control points/landmarks that are located on the boundary (a closed curve in 2D cases or a surface in 3D cases) of the object of interest. As an example, Fig. 5.2(a) shows the distribution of the landmarks on the boundary of the Medial Gastrocnemius (MG) muscle, which were considered in the experiments of calf muscle segmentation (see section 5.4.2). Let x_i ($i \in \mathcal{V}$), a 3-dimensional vector, denote the 3d position of landmark i and $\mathbf{x} = (x_i)_{i \in \mathcal{V}}$ denote the position of all the landmarks which parameterize the surface.

We consider training data $\mathcal{M}_{\text{train}}$ which consist of a set of M shapes, *i.e.*, $\mathcal{M}_{\text{train}} = \{\mathbf{x}^{(m)}\}_{m\in\{1,2,\dots,M\}}$, to learn a prior probability distribution on the configuration of the 3D shape model. As have been presented in section 5.1.1, we aim to achieve a pose-invariant prior model. Thus, we do not register all the surfaces into a reference space. However, we assume that correspondences have been determined for the landmarks among the samples of the training set. Based on such training data, we propose to learn statistics on measurements that are invariant with respect to translations, rotations and scales and can be encoded within small cliques of an MRF model.

Let us consider a clique c ($c \subseteq \mathcal{V}$ and $|c| \geq 3$) of landmarks, we enumerate all the pairs $\mathcal{P}_c = \{(i,j)|i,j \in c \text{ and } i < j\}$ of points. Let $d_{ij} = \|x_i - x_j\|$ denote the Euclidean distance between points i and j ($(i,j) \in \mathcal{P}_c$). We obtain the relative distance $\hat{d}_{ij}$ by normalizing the distance d_{ij} over the sum of the distances between the pairs of points involved in clique c, *i.e.*,

$$\hat{d}_{ij} = \frac{d_{ij}}{\sum_{(i,j)\in\mathcal{P}_c} d_{ij}} \tag{5.1}$$

Since for clique c, any relative distance $\hat{d}_{ij}$ is a linear combination of the others (*i.e.*, $\sum_{(i,j)\in\mathcal{P}_c} \hat{d}_{ij} = 1$), we store all the relative distances, except one in a vector $\hat{\mathbf{d}}_c$, *i.e.*,

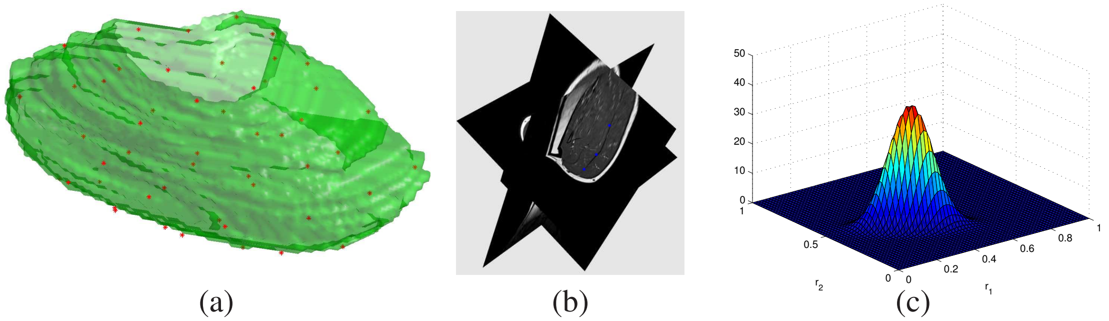
$$\hat{\mathbf{d}}_c = (\hat{d}_{ij})_{(i,j)\in\bar{\mathcal{P}}_c} \tag{5.2}$$

where $\bar{\mathcal{P}}_c$ contains the pairs that are involved in the vector $\hat{\mathbf{d}}_c$. For the purpose of clarity, let us consider third-order cliques (*i.e.*, $|c| = 3$) as an example, which is used in our knowledge-based segmentation formulation that will be presented in section 5.2.3. In a third-order clique $c = \{i,j,k\}$ ($i,j,k \in \mathcal{V}$ and $i < j < k$), the corresponding three points compose a triangle Δ_{ijk} and $\hat{\mathbf{d}}_c$ denotes the relative lengths $(\hat{d}_{ij}, \hat{d}_{jk})$ of the sides (i,j) and (j,k), *i.e.*,

$$\hat{\mathbf{d}}_c = \left(\frac{d_{ij}}{d_{ij}+d_{jk}+d_{ki}}, \frac{d_{jk}}{d_{ij}+d_{jk}+d_{ki}}\right) \tag{5.3}$$

The statistics on $\hat{\mathbf{d}}_c$ are learned from the training data. We can model its distribution $\psi_c(\hat{\mathbf{d}}_c)$ using standard probabilistic models such as Multivariate Gaussian Distributions, Gaussian Mixtures, Parzen-Windows. Finally, we get the higher-order shape model $\mathcal{S} = (\mathcal{V}, \mathcal{C}, \{\psi_c(\cdot)\}_{c\in\mathcal{C}})$, where $\mathcal{V}$ and $\mathcal{C}$ determine the topology of the model while $\{\psi_c(\cdot)\}_{c\in\mathcal{C}}$ characterizes the statistical geometric constraints between the points contained in each clique $c \in \mathcal{C}$. In the case where third-order cliques are used, $\mathcal{C}$ is defined as $\mathcal{C} = \{\{i,j,k\}|i,j,k \in \mathcal{V} \text{ and } i < j < k\}$. Such statistical constraints can be easily encoded in a higher-order MRF with a clique set that includes $\mathcal{C}$, which results in a prior probability on the 3D configuration of the shape model as follows:

$$p(\mathbf{x}) \propto \prod_{c\in\mathcal{C}} \psi_c(\hat{\mathbf{d}}_c(x_c)) \tag{5.4}$$



(a) (b) (c)

Figure 5.2: Shape Model for Medial Gastrocnemius Muscle. (a) Distribution of the landmarks on the muscle boundary. (b) Two perpendicular slices with a triplet of landmarks (the blue asterisks). (c) Learned Gaussian distribution on $\hat{\mathbf{d}}_c$ for the triplet shown in (b).

where $\hat{\mathbf{d}}_c(x_c)$ denotes the mapping from the 3D positions x_c of the three points contained in the clique c to the relative distance vector $\hat{\mathbf{d}}_c$.

5.2.2 Landmark Candidate Detection

As presented in section 5.1.1, we aim at developing a one-shot optimization approach for the segmentation of challenging medical image data such as MRI data of calf muscles. In order to explore image support through feature vectors and to avoid a prohibitive computational complexity, we perform landmark detections to find a set of possible correspondences (referred to as "candidates") in the observed image for each landmark i ($i \in \mathcal{V}$) in the 3D shape model. To this end, we first learn a classifier for each landmark, and then compute a score for each possible location, and finally select the L positions that have the best scores to compose the candidate set for the landmark.

There are various standard classifiers such as *Randomized Forests* [Breiman 2001], *Boosting* algorithms [Schapire 1990, Freund & Schapire 1997, Schapire 2001] and *Support Vector Machines (SVMs)* [Boser *et al.* 1992, Cortes & Vapnik 1995, Muller *et al.* 2001]. In this work, we employ Randomized Forests to perform the classification. However, our method is modular with respect to the classifier and other classifiers can also be considered. *Randomized forests* [Breiman 2001] were developed based on previous works on "Bagging" (*i.e.*, *Bootstrap aggregating*) and the random selection of features (*e.g.*, [Breiman 1996, Amit & Geman 1997, Ho 1998]). They provide a powerful tool for classifications and has been successfully applied in various computer vision problems, such as object recognition [Lepetit & Fua 2006], image classification [Bosch *et al.* 2007], object segmentation via graph cuts [Winn & Shotton 2006, Schroff *et al.* 2008] and facade segmentation/parsing using procedural shape prior [Simon *et al.* 2011].

A *randomized forest* is composed of a set of T random decision trees. In the decision trees, an internal node consists of a random test on an input feature vector. When a feature vector is presented to a tree, it follows a specific path down to a leaf node. At each step of this path, the direction (left or right) is determined by the binary result of the random test (corresponding to the internal node) applied to the input vector. A leaf node stores a histogram $h = (h_1, \ldots, h_W)$ (W is the number of classes), which is obtained during the training phase by counting the number of labeled feature vectors that arrive at this leaf. During the testing phase, an unlabeled feature vector is dropped in each decision tree τ and eventually reaches the leaf l_τ. The normalized histogram of l_τ provides a probability estimation for the feature vector belonging to each class w:

$$P(w|l_\tau) = \frac{h_w}{\sum_i h_i} \tag{5.5}$$

Finally, the probabilities of all the trees are averaged to obtain the probability over the forest:

$$P(w|(l_1, \ldots, l_T)) = \frac{1}{T} \sum_\tau P(w|l_\tau) \tag{5.6}$$

We consider all the voxels in a 3D volume as possible locations of the landmarks. Each voxel is associated with a feature vector that is used as the input for classifiers. Different features can be considered in randomized forests towards achieving a high-quality detection. Image patches centered at each voxel are certainly the most straightforward features to use. A more sophisticated one consists of a series of 3D Gabor features [Bernardino & Santos-Victor 2006] with different scale, rotation parameters, which can well capture the local image structure information. Furthermore, these parameters can be sampled using the method proposed in [Kokkinos & Yuille 2008] so that scaling/rotation of the image becomes a translation of these parameters, and then the Fourier Transform Modulus (FTM) of the filter output can be estimated to eliminate variations due to these translations (because the FTM is translation invariant). Due to the symmetry of FTM, it is enough to consider only half of the FTM domain by removing the redundant coefficients, which results in a scale and rotation invariant feature vector. Fig. 5.3 shows the detected candidate results for four landmarks at different locations on a testing muscle data.

5.2.3 Higher-order MRF Segmentation Formulation

The shape model, together with the evidence from the image support, is formulated within a higher-order MRF towards image segmentation. To this end, let $\mathcal{G} = (\mathcal{V}, \mathcal{C})$ denote a hypergraph[2] with a node set $\mathcal{V}$ and a clique set $\mathcal{C}$. We associate each landmark to a node

[2] We reuse the notation $\mathcal{V}$ and $\mathcal{C}$ to denote the node set and the clique set of the hypergraph, respectively, due to one-to-one mappings between the nodes/cliques of the hypergraph and the landmarks/cliques of the

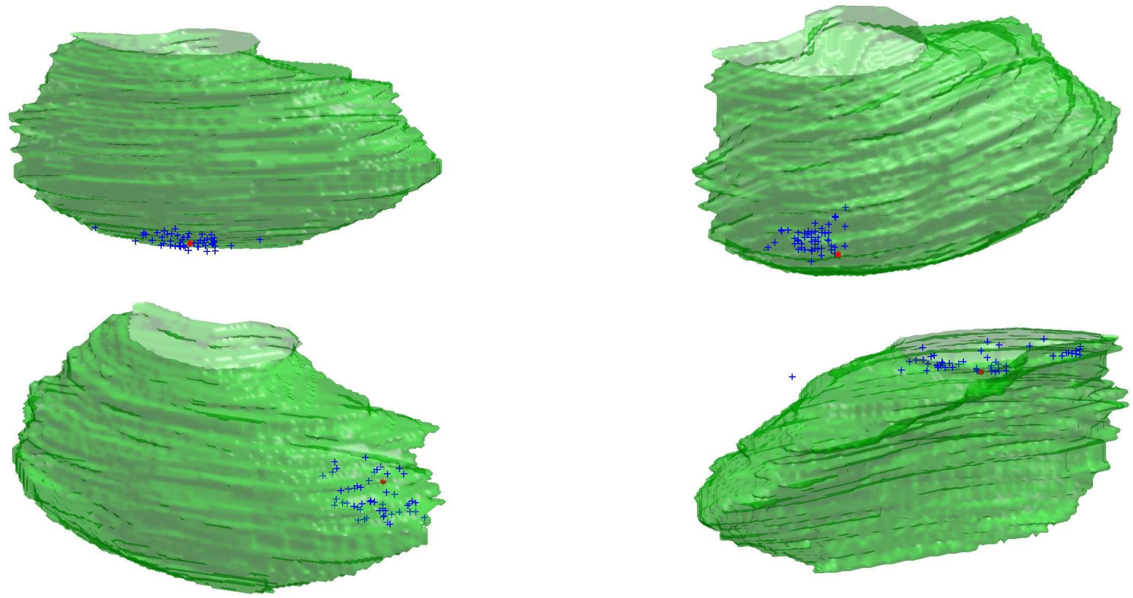


Figure 5.3: Landmark Detection Results. The red hexagram represents the ground truth while the blue plus signs represent the 50 candidates that have the best scores during the detection. The reference segmentation surface is provided to visually measure the distance between candidates and the ground truth.

i ($i \in \mathcal{V}$) in the hypergraph, and the latent variable X_i corresponding to the node i is a 3-dimensional vector that denotes the 3D position of the associated landmark. The candidate set of each variable is denoted by $\mathcal{X}_i$ ($i \in \mathcal{V}$), which consists of the detected landmark candidates (see section 5.2.2). Thus the Cartesian product $\mathcal{X} = \prod_{i \in \mathcal{V}} \mathcal{X}_i$ denotes the candidate set of the configuration $\mathbf{x} = (x_i)_{i \in \mathcal{V}}$ of the MRF model. In this work, we use the pose-invariant shape prior of third order (see section 5.2.1 for the definition of the shape prior). In order to introduce such a prior into the MRF formulation, we associate a triplet of landmarks to a third-order clique c and use the potential function of the clique c to encode the statistical spatial constraints between the three landmarks. Finally, the segmentation problem is transformed into estimating the optimal positions of the landmarks, *i.e.*, the optimal configuration $\mathbf{x}^{\text{opt}}$ of the higher-order MRF, which is formulated as a minimization of the MRF energy $E(\mathbf{x})$:

$$\mathbf{x}^{\text{opt}} = \arg\min_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}) \tag{5.7}$$

shape model.

The energy of MRF is defined as a sum of singleton potentials $U_i(x_i)$ $(i \in \mathcal{V})$ and third-order potentials $U_c(x_c)$ $(c \in \mathcal{C})$, *i.e.*,

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} U_i(x_i) + \sum_{c \in \mathcal{C}} H_c(x_c) \tag{5.8}$$

where x_c denotes the configuration $(x_i)_{i \in c}$ of clique c. The singleton potentials and third-order clique potentials are presented below.

The singleton potential $U_i(x_i)$ $(i \in \mathcal{V})$ consists of the negative log-likelihood which imposes penalty for the landmark i being located at position x_i in image $\mathbf{I}$, *i.e.*,

$$U_i(x_i) = -\log p(\mathbf{I}|x_i) \tag{5.9}$$

$p(\mathbf{I}|x_i)$ is defined using the classifier's output probability value for landmark i being located at x_i.

The higher-order clique potential $U_c(x_c)$ $(c \in \mathcal{C})$ encodes the statistic geometry constraints between the triplet c of points and is defined as:

$$U_c(x_c) = -\alpha \cdot \log \psi_c(\hat{\mathbf{d}}_c(x_c)) \tag{5.10}$$

where $\alpha > 0$ is a weight coefficient, $\hat{\mathbf{d}}_c(x_c)$ denotes the mapping from the position of the triplet c to the 2-dimensional relative lengths of the sides, and $\psi_c(\cdot)$ denotes the learned distribution on the relative lengths (see section 5.2.1).

5.3 3D Landmark Model Inference from Monocular 2D Images

5.3.1 Probabilistic 3D-2D Inference Framework

We consider a point-distribution shape model composed of a set $\mathcal{V}$ of landmarks located on the surface of the 3D object of interest. Let latent variable $X_i = (X_i^{(3)}, X_i^{(2)})$ denote the 3D and 2D positions of a landmark i $(i \in \mathcal{V})$. More specifically, $X_i^{(3)}$ and $X_i^{(2)}$, 3-dimensional and 2-dimensional vectors respectively, denote the 3D position of landmark i in the model space and the 2D position in the observed image $\mathbf{I}$. Each variable X_i takes a value x_i from its possible configuration set $\mathcal{X}_i = \mathcal{X}_i^{(3)} \times \mathcal{X}_i^{(2)}$, where $\mathcal{X}_i^{(3)}$ and $\mathcal{X}_i^{(2)}$ denote the 3D and 2D position candidate sets, respectively. Due to the fact that landmarks may be invisible, we also introduce a visibility variable O_i for landmark i [Sudderth *et al.* 2004a]. $O_i = 1$ when the landmark is visible in the 2D image space, and $O_i = 0$ otherwise.

Given the observed image $\mathbf{I}$, the estimation of the 3D-2D positions and the visibility of the landmarks is formulated as a maximization of the posterior probability of $(\mathbf{X}, \mathbf{O}) =$

$((X_i)_{i\in\mathcal{V}}, (O_i)_{i\in\mathcal{V}})$ over their domains $\mathcal{X} = \prod_{i\in\mathcal{V}} \mathcal{X}_i$ and $\mathcal{O} = \{0, 1\}^{|\mathcal{V}|}$:

$$(\mathbf{x}, \mathbf{o})^{\text{opt}} = \underset{(\mathbf{x},\mathbf{o})\in\mathcal{X}\times\mathcal{O}}{\arg\max}\; p(\mathbf{x}, \mathbf{o}|\mathbf{I}) \tag{5.11}$$

The posterior probability $p(\mathbf{x}, \mathbf{o}|\mathbf{I})$ is:

$$\begin{aligned} p(\mathbf{x}, \mathbf{o}|\mathbf{I}) &= p(\mathbf{x}, \mathbf{o}, \mathbf{I})/p(\mathbf{I}) \\ &\propto p(\mathbf{x}, \mathbf{o}, \mathbf{I}) \\ &= p(\mathbf{I}|\mathbf{x}, \mathbf{o}) \cdot p(\mathbf{x}, \mathbf{o}) \\ &= p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{o}) \cdot p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o}) \cdot p(\mathbf{o}|\mathbf{x}^{(3)}) \cdot p(\mathbf{x}^{(3)}) \\ &= \underbrace{p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})}_{\text{Image Likelihood}} \cdot \underbrace{p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o})}_{\text{Projection Prior}} \cdot \underbrace{p(\mathbf{o})}_{\text{Visibility Prior}} \cdot \underbrace{p(\mathbf{x}^{(3)})}_{\text{3D Model Prior}} \end{aligned} \tag{5.12}$$

where $p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})$ encodes the likelihood of the observed image given the 2D position configurations $\mathbf{x}^{(2)}$ and the visibility states $\mathbf{o}$ of the landmarks, $p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o})$ encodes the projection prior from the 3D configuration $\mathbf{x}^{(3)}$ to the 2D configuration of the landmarks, $p(\mathbf{o})$ denotes the visibility prior on the landmarks, and $p(\mathbf{x}^{(3)})$ denotes the prior on the 3D configurations of the landmarks.

Note that this probabilistic formulation can be directly applied to the estimation of 3D (or 2D) configuration of the landmarks given 2D (or 3D) configuration, simply by instantiating the variables whose configurations are known.

5.3.2 Definitions of the Probability Terms

In this section, let us elaborate all the probability terms which are involved in the posterior probability $p(\mathbf{x}, \mathbf{o}|\mathbf{I})$ (see Eq. 5.12).

Image Likelihood

The image likelihood $p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})$ measures the occurrence probability of the observed image $\mathbf{I}$, given the 2D position configurations $\mathbf{x}^{(2)}$ and the visibility states $\mathbf{o}$ of the landmarks. If we assume, without loss of generality, that the landmarks are independent in terms of appearance, then we can define $p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})$ as follows:

$$p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o}) \propto \prod_{i\in\mathcal{V}} p(\mathbf{I}|x_i^{(2)}, o_i) \tag{5.13}$$

Regarding $p(\mathbf{I}|x_i^{(2)}, o_i)$, there are two possible cases as follows:

1. When $O_i = 1$, the landmark's position is informative. In such a case, $p(\mathbf{I}|x_i^{(2)}, o_i)$ denotes the likelihood of the observed image given that landmark i is located at position $x_i^{(2)}$, which can be defined using the output of a classifier such as Randomized Forest [Breiman 2001].

2. When $O_i = 0$, the landmark's position is not informative. In this case, $p(\mathbf{I}|x_i^{(2)}, o_i)$ denotes a uniform distribution, thus we assume that $p(\mathbf{I}|x_i^{(2)}, o_i) = \hat{p}$ (constant).

Projection Prior

The projection prior $p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o})$ measures the occurrence possibility of the 2D positions $\mathbf{x}^{(2)}$ of the landmarks when the 3D positions $\mathbf{x}^{(3)}$ and the visibility states $\mathbf{o}$ are given, which is modeled using Gibbs distribution:

$$p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o}) \propto \exp\{-\frac{f(\mathbf{x}, \mathbf{o})}{T}\} \tag{5.14}$$

where T is temperature, and the energy function $f(\mathbf{x}, \mathbf{o})$ encodes inconsistency between the 3D and 2D configurations of the landmarks taking the visibility states into account (the smaller $f(\mathbf{x}, \mathbf{o})$ is, the better is the correspondence between $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(2)}$).

Without loss of generality, we use the weak-perspective camera configuration [Alter 1994] to model the projection from 3D points to 2D points[3]. Let us first consider a triplet $t \in \mathcal{T} = \{t|t \subseteq \mathcal{V} \text{ and } |t| = 3\}$ of landmarks that are all visible. Their 3D-2D positions x_t determine at most two projection mappings $\mathbf{P}_{x_t}^{(s)}$ $(s \in \{1, 2\})$ [Alter 1994, Fischler & Bolles 1981] corresponding to two reflective symmetric camera configurations. Then for any additional visible point i, we can measure the error $e_{x_t}(x_i)$ between its 2D position $x_i^{(2)}$ and the value obtained by projecting its 3D position $x_i^{(3)}$, *i.e.*:

$$e_{x_t}(x_i) = \min_{s \in \{1,2\}} \left\| \mathbf{P}_{x_t}^{(s)}(x_i^{(3)}) - x_i^{(2)} \right\| \tag{5.15}$$

where $\|\cdot\|$ denotes the Euclidean norm, and between the two feasible projections we consider the most prominent one with respect to the considered 2D configuration [Alter 1994]. On the contrary, if one or more of these four points are invisible, we set a constant energy $\hat{E}$ as the projection error $e_{x_t}(x_i)$, which can be understood as an upper bound of the average projection error which is allowed between four points. Therefore, we define the error function $e_{x_t,o_t}(x_i, o_i)$ by taking the visibility states into account as:

$$e_{x_t,o_t}(x_i, o_i) = w_t \cdot \begin{cases} e_{x_t}(x_i) & \text{if } o_j = 1, \forall j \in t \cup \{i\} \\ \hat{E} & \text{otherwise} \end{cases} \tag{5.16}$$

[3] In the proposed framework, the weak-perspective camera model can be easily replaced by other camera models such as the perspective model.

where w_t is a confidence weight for the error measure obtained under the mapping determined by the positions of the points in clique t, which will be presented later in this section. And then, the 3D-2D consistency between a quadruplet c of landmarks consists of the sum of the errors which are determined by taking all possible combinations of triplets within the quadruplet and evaluating the projection error on the remaining point, which can be formulated mathematically as follows:

$$e(x_c, o_c) = \sum_{t \subset c} e_{x_t, o_t}(x_{c \setminus t}, o_{c \setminus t}) \tag{5.17}$$

Finally, we define the energy function $f(\mathbf{x}, \mathbf{o})$ as the sum of $e(x_c, o_c)$ over all the quadruplet, *i.e.*:

$$f(\mathbf{x}, \mathbf{o}) = \sum_{c \in \mathcal{C}} e(x_c, o_c) \tag{5.18}$$

where $\mathcal{C} = \{c | c \subseteq \mathcal{V} \text{ and } |c| = 4\}$ denotes the set of all quadruplets. Last, we should note that we can further combine other cues in this projection prior, such as regional texture similarity.

Robust Confidence Weight

Since the projection matrix estimation is unstable when considering triplets of 3D points that are nearly collinear [Alter 1994], we introduce a confidence weight w_t to modulate the error contribution of each triplet of points. For a triangle $\Delta_{x_t^{(3)}}$ consisting of a triplet t of points with 3D positions $x_t^{(3)}$, we define the non-collinear coefficient $\text{NC}(x_t^{(3)})$ using the square root of its area $\text{Area}(\Delta_{x_t^{(3)}})$ and its perimeter $\text{Perim}(\Delta_{x_t^{(3)}})$ as follows:

$$\text{NC}(x_t^{(3)}) = \frac{2 \times 3^{\frac{3}{4}} \times \text{Area}^{\frac{1}{2}}(\Delta_{x_t^{(3)}})}{\text{Perim}(\Delta_{x_t^{(3)}})} \tag{5.19}$$

We can observe that $\text{NC}(x_t^{(3)}) = 1$ for an equilateral triangle and $\text{NC}(x_t^{(3)}) = 0$ when the three points are collinear. Then we learn the confidence weight w_t by averaging the non-collinear coefficients for each triplet t over the training data:

$$w_t = \frac{1}{M} \sum_{m=1}^{M} \text{NC}(x_{t,m}^{(3)}) \tag{5.20}$$

where M denotes the number of training samples.

Specification of the Projection Error

Regarding the computation of $e_{x_t}(x_i)$, we use the efficient method proposed in [Alter 1994] to compute directly the projection of a 3D point under the projection determined by a triplet of corresponding 3D-2D points without calculating the projection mapping. We refer readers to [Alter 1994] for more details.

Collinear triplets of points lead to degenerate configurations from which we cannot obtain a solution for the projection mapping. In this case, the corresponding error term $e_{x_t}(x_i)$ in Eq. 5.15 is not well-defined. To deal with this, we consider two different scenarios: (i) When we have a prior knowledge that the 3D positions of a triplet t of points have to be collinear, we simply ignore the corresponding error measure by defining $e_{x_t}(x_i) = 0$ (this is consistent with the confidence weight defined in Eq. 5.20, *i.e.*, $w_t = 0$ leads to zero contribution to $f(\mathbf{x})$); (ii) Otherwise, we define $e_{x_t}(x_i) = +\infty$ if $\mathbf{x}_t^{(3)}$ are collinear so that the final solution of $\mathbf{x}_t^{(3)}$ cannot be exactly collinear. By doing so, the term $e_{x_t}(x_i)$ is well-defined for all the cases. For the sake of clarity, hereafter, we assume that the definition of $e_{x_t}(x_i)$ in Eq. 5.15 implicitly includes the definition in the degenerate case.

Visibility Prior

We introduce the visibility variable $\mathbf{O}$ to achieve a more precise modeling of the 3D-2D estimation, due to the fact that a landmark can be invisible. The notion of "invisibility" encodes occlusions and self-occlusions in the 3D space, as well as misdetection due to insufficient image support or classification failure.

To better understand such a notion of "invisibility", one can consider that the visibility of landmark is with respect to the landmark detector. Let us elaborate this in the considered problem. The inference process is performed by considering, for each landmark i, a number of 2D positions which lead to the highest probabilities $p(\mathbf{I}|x_i^{(2)})$ towards composing the set of plausible solutions for $\mathcal{X}_i^{(2)}$, expecting that at least one candidate is (or close to) the true position. However, because of erroneous detection or occlusions, it is possible that all the candidates are far from the ground truth. In such a context, we define the notion of "visibility" as whether the true 2D correspondence of the landmark is captured by the candidate set. More specifically, $O_i = 1$ means that at least one candidate in $\mathcal{X}_i^{(2)}$ is close to the ground truth, and $O_i = 0$ stands for the opposite case.

The prior probability $p(\mathbf{o})$ is defined as follows:

$$p(\mathbf{o}) = \prod_{i \in \mathcal{V}} p(o_i) \tag{5.21}$$

where $p(o_i)$ denotes the prior probability of the visibility of each individual landmark i and is modeled as a Bernoulli distribution $\mathrm{Bern}(o_i|\mu_i)$ with parameter $\mu_i = \Pr(O_i = 1)$.

In practice, it is usually reasonable to assume the same parameter $\mu > 0.5$ for all the landmarks [Torresani *et al.* 2008].

3D Model Prior

We adopt the pose-invariant shape model introduced in section 5.2.1 that can capture the inherent variability of the class of objects from a reasonable small training set and can be easily modeled within MRFs. Thank to the invariance under similarity transformation, no registration between surfaces is required during the learning stages and we only assume that correspondences have been determined for the landmarks among the samples of the training set. Due to the fact that the projection prior (section 5.3.2) involves quadruplets of points, we instantiate the generic shape model in section 5.2.1 using fourth-order cliques (*i.e.*, $|c| = 4$). Similar to Eq. 5.4, the prior probability on the 3D positions of the landmarks is defined as follows:

$$p(\mathbf{x}^{(3)}) \propto \prod_{c \in \mathcal{C}} \psi_c(\hat{\mathbf{d}}_c(x_c^{(3)})) \tag{5.22}$$

5.3.3 Higher-order MRF Formulation

The data likelihood, the 3D-2D consistency, the visibility prior and the 3D shape model, presented in section 5.3.2, can be naturally encoded within a higher-order MRF model where latent variables are to be inferred through an energy minimization. In this perspective, the negative logarithm of the posterior probability (Eq. 5.12) is factorized into the potentials of the MRF and constitutes the MRF energy.

To this end, we use a node to model a landmark i ($i \in \mathcal{V}$) with its latent 3D-2D position X_i and its visibility o_i. Actually, we can use a single random variable[4] to encode X_i and o_i compactly by simply defining a special label "occ" within 2D position candidate set $\mathcal{X}_i^{(2)}$ such that:

$$x_i = \begin{cases} (x_i^{(3)}, x_i^{(2)}) & \text{if } O_i = 1 \\ (x_i^{(3)}, \text{occ}) & \text{if } O_i = 0 \end{cases} \tag{5.23}$$

This compact representation is valid because the 2D position $x_i^{(2)}$ is meaningless when the landmark i is occluded (*i.e.*, when $O_i = 0$, the image likelihood $p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})$ and the projection prior $p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o})$ are constant with respect to $x_i^{(2)}$.).

In order to factorize the potential functions, we use a fourth-order clique to model a quadruplet c of landmarks. Due to the bijective mappings between nodes and landmarks and between fourth-order cliques and quadruplets, we reuse $\mathcal{V}$ and $\mathcal{C}$ to denote the node

[4] In order to reduce the number of symbols used, we reuse X_i to denote this new random variable. Accordingly, we reuse x_i, $\mathcal{X}_i$ and the other related notations.

set and the clique set which determine the topology of the MRF. The 3D and 2D positions of the landmarks are estimated through the minimization of the MRF energy $E(\mathbf{x})$:

$$\mathbf{x}^{\text{opt}} = \arg\min_{\mathbf{x}\in\mathcal{X}} E(\mathbf{x}) \tag{5.24}$$

Here, the energy of the MRF is defined as the negative logarithm of the posterior probability in Eq. 5.12 (up to an additive constant) and can be factorized into the following form:

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}} U_i(x_i) + \sum_{c\in\mathcal{C}} H_c(x_c) \tag{5.25}$$

where x_c denotes the configuration $(x_i)_{i\in c}$ of clique c.

Algorithm 5.1 Decompose A Factor Graph into Factor Trees

```
Input: Factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{C})$ with the node set $\mathcal{V}$ and the factor set $\mathcal{C}$
Output: A set of factor trees $\Gamma_{\mathcal{G}} = \{\mathcal{G}_s = (\mathcal{V}_s, \mathcal{C}_s)\}_{s\in\mathcal{S}}$
  $\Gamma_{\mathcal{G}} \leftarrow \emptyset$
  while $\mathcal{C} \neq \emptyset$ do
    Get an factor $c$ from the factor set $\mathcal{C}$
    $\mathcal{V}_s \leftarrow \mathcal{N}_c$ {$\mathcal{N}_c$ denotes the set of neighbor nodes of the factor $c$ in $\mathcal{G}$}
    $\mathcal{C}_s \leftarrow \{c\}$
    $\mathcal{C}_{\text{ext}} \leftarrow \{c\}$
    while $\mathcal{C}_{\text{ext}} \neq \emptyset$ do
      Get an factor $c'$ from the factor set $\mathcal{C}_{\text{ext}}$
      $\mathcal{C}_{\text{ext}} \leftarrow \mathcal{C}_{\text{ext}} \setminus \{c'\}$
      $\mathcal{V}_{\text{ext}} \leftarrow \mathcal{N}_{c'}$
      for all $v' \in \mathcal{V}_{\text{ext}}$ do
        for all $\hat{c} \in \mathcal{N}_{v'} \setminus \{c'\}$ {$\mathcal{N}_{v'}$ denotes the set of neighbor factors of the node $v'$ in
        $\mathcal{G}$} do
          if the graph $(\mathcal{V}_s \cup \mathcal{N}_{\hat{c}}, \mathcal{C}_s \cup \{\hat{c}\})$ has no loop then
            $\mathcal{V}_s \leftarrow \mathcal{V}_s \cup \mathcal{N}_{\hat{c}}$
            $\mathcal{C}_s \leftarrow \mathcal{C}_s \cup \{\hat{c}\}$
            $\mathcal{C}_{\text{ext}} \leftarrow \mathcal{C}_{\text{ext}} \cup \{\hat{c}\}$
          end if
        end for
      end for
    end while
    $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{C}_s)$
    $\mathcal{C} \leftarrow \mathcal{C} \setminus \mathcal{C}_s$
    $\Gamma_{\mathcal{G}} \leftarrow \Gamma_{\mathcal{G}} \cup \{\mathcal{G}_s\}$
  end while
```

Singleton potential

The singleton potential $U_i(x_i)$ ($i \in \mathcal{V}$) encodes the data likelihood (see section 5.3.2) and the visibility prior (see section 5.3.2). After taking the negative logarithm, we obtain its definition as follows:

$$U_i(x_i) = \begin{cases} -\log p(\mathbf{I}|x_i^{(2)}) & \text{if } x_i^{(2)} \neq \text{“occ”} \\ \lambda_1 & \text{if } x_i^{(2)} = \text{“occ”} \end{cases} \tag{5.26}$$

where λ_1 is a constant coefficient.

Higher-order clique potential

The higher-order clique potential $H_c(x_c)$ ($c \in \mathcal{C}$) is defined as follows:

$$H_c(x_c) = \lambda_2 \cdot H_c^{(1)}(x_c) + \lambda_3 \cdot H_c^{(2)}(x_c) \tag{5.27}$$

where $\lambda_2 > 0$ and $\lambda_3 > 0$ are two balancing constants, $H_c^{(1)}(x_c)$ encodes the 3D statistic geometry constraints implied by the shape prior on the 3D configuration of the landmarks, and $H_c^{(2)}(x_c)$ encodes the 3D-2D projection prior:

$$\begin{cases} H_c^{(1)}(x_c) = -\log \psi_c(\hat{\mathbf{d}}_c(x_c^{(3)})) \\ H_c^{(2)}(x_c) = e(x_c, o_c(x_c)) \end{cases} \tag{5.28}$$

where $o_c(x_c)$ denotes the binary visibility values that are recovered from x_c using Eq. 5.23, and the definitions of $e(x_c, o_c)$ and $\psi_c(\alpha(x_c^{(3)}))$ have been presented in section 5.3.1.

5.4 Experimental Results

The optimization of the MRF models for both problems requires a higher-order MRF-MAP inference algorithm. We present first in section 5.4.1 the optimization approach that was used in the experiments and then show the experimental results in section 5.4.2 and section 5.4.3.

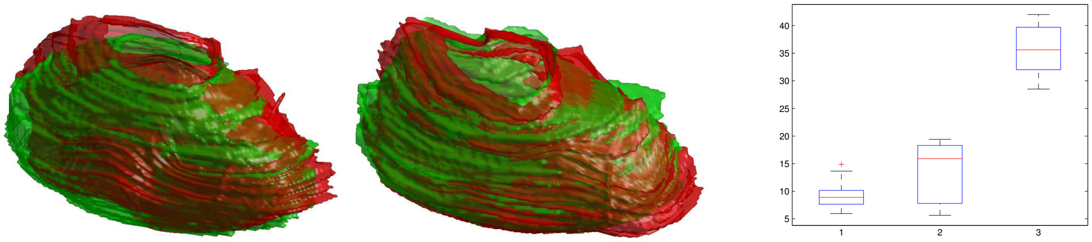
5.4.1 Higher-order MRF Optimization via Dual-decomposition

In section 2.2.3, we have reviewed the *dual-decomposition* MRF optimization framework [Bertsekas 1999, Komodakis *et al.* 2007a], which has also been applied in solving higher-order MRFs [Komodakis & Paragios 2009] and other specific problems such as graph matching [Torresani *et al.* 2008] and joint segmentation and appearance histogram models optimization [Vicente *et al.* 2009]. We also have adopted such a framework for solving high-order graph matching problems as shown in chapter 4. Motivated by the advantages of the dual-decomposition framework in terms of flexibility, generality and convergence property (see section 2.2.3) and by promising performance achieved by the dual-decomposition-based methods developed in those previous works, we chose to adopt such an optimization methodology to perform the inference in our higher-order MRF models as well. However, due to the fact that the higher-order potentials contained in the MRFs are not pattern-based, the techniques proposed in [Komodakis & Paragios 2009] cannot be directly used. Thanks to the flexibility and generality of dual-decomposition, one can resort to other kinds of decompositions in order to optimize the energy of the MRF. Tree decompositions have been widely used in the literature to develop successful MRF-MAP inference algorithms (*e.g.*, [Wainwright *et al.* 2005, Kolmogorov 2006, Komodakis *et al.* 2007a]), due to the fact that the inference in a tree can be exactly done in polynomial computational time (see section 2.2.2). Using *factor trees* (see section 2.1.4), many properties and algorithms of usual trees can be generalized to higher-order cases, such as the min-sum belief propagation (see Algorithm 2.2).

Based on these observations, we adopt the dual-decomposition optimization framework and decompose the original problem into a set of sub-problems each of which corresponds to a factor-tree. More concretely, we represent an MRF as a *factor graph* (see section 2.1.4). Let MRF$^{\mathcal{G}}$ denote the original MRF model whose topology is defined by the factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{C})$, $\mathbf{U}^{\mathcal{G}} = \{U_i(\cdot)\}_{i\in\mathcal{V}}$ denotes the singleton potentials defined on the node set $\mathcal{V}$, and $\mathbf{H}^{\mathcal{G}} = \{H_c(\cdot)\}_{c\in\mathcal{C}}$ denotes the clique potentials defined on the factor set $\mathcal{C}$. We decompose the original hypergraph $\mathcal{G}$ into a set of factor trees, which are denoted by $\{\mathcal{G}_s = (\mathcal{V}_s, \mathcal{C}_s)\}_{s\in\mathcal{S}}$, such that $\mathcal{V} = \cup_{s\in\mathcal{S}}\mathcal{V}_s$, $\mathcal{C} = \cup_{s\in\mathcal{S}}\mathcal{C}_s$ and any higher-order factor in $\mathcal{G}$ appears in one and only one factor tree. This process can be easily done using the algorithm described in Algorithm 5.1. The potentials of the MRFs corresponding to the sub-problems, denoted by $\{\text{MRF}^{\mathcal{G}_s}\}_{s\in\mathcal{S}}$, are obtained by decomposing the potentials of the original MRF into the sub-hypergraphs such that $\mathbf{U}^{\mathcal{G}} = \sum_{s\in\mathcal{S}}\mathbf{U}^{\mathcal{G}_s}$ and $\mathbf{H}^{\mathcal{G}} = \sum_{s\in\mathcal{S}}\mathbf{H}^{\mathcal{G}_s}$. This can be achieved simply by setting $U_i^{\mathcal{G}_s} = \frac{U_i^{\mathcal{G}}}{|\{s|i\in\mathcal{V}_s\}|}$ and $H_c^{\mathcal{G}_s} = \frac{H_c^{\mathcal{G}}}{|\{s|c\in\mathcal{C}_s\}|}$. Max-product belief propagation algorithm (see Algorithm 2.2) is employed to exactly and efficiently perform the inference in each subproblem and the solutions of the sub-problems are combined using projected subgradient method (like [Komodakis *et al.* 2007a, Torresani *et al.* 2008, Komodakis & Paragios 2009]) to solve the Lagrangian dual so as to obtain the solution of the original problem.

5.4.2 Results on Knowledge-based 3D Segmentation

We used the data set that was previously used in [Essafi *et al.* 2009a] to validate the proposed method. This data set consists of 25 3D MRI subjects whose calf part was captured.



(a)

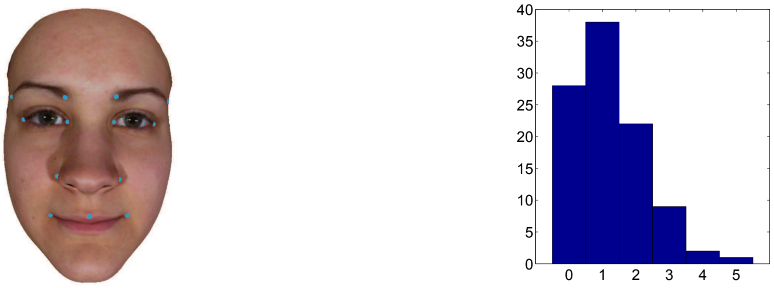
(b)

Figure 5.4: Experimental Results for Muscle Segmentation. (a) Surface reconstruction results (green: reference segmentation. red: reconstruction result). (b) Boxplots of the average landmark error measure in voxel (1. our method. 2. method in [Essafi *et al.* 2009a]. 3. standard ASM method.). On each box, the central mark in red is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points.

The voxel spacing is of $0.7812 \times 0.7812 \times 4$ mm and each volume consists of 90 slices of 4mm thickness acquired with a 1.5 T Siemens scanner. Standard of reference was available, consisting of annotations provided by experts for the Medial Gastrocnemius (MG) muscle.

We performed a leave-one-out cross validation on the whole data set. For comparison purpose, we considered as alternative segmentation methods[5] the ones presented in [Essafi *et al.* 2009a]. We present in Fig. 5.4(a) the surface reconstruction results using the estimated position of the landmarks and thin plate spline (TPS) [Bookstein 1989], while in Fig. 5.4(b) the average distance between the real landmark position and the one estimated from our algorithm, and the ones reported in [Essafi *et al.* 2009a] including the one obtained using standard active shape models. In comparison to [Essafi *et al.* 2009a], considered as state-of-the-art, our approach leads to an average reduction of the landmark location error by a factor 2. The analysis of the results shows that the proposed prior and the inference using higher-order graphs globally perform well while the main limitation is introduced from the landmark candidate detection process. Since the method establishes correspondences between the model and the detected landmarks, in the absence of meaningful candidates the method fails to recover optimally the global shape. Regarding computational complexity, the method is linear with respect to the number of higher-order cliques and cubic with respect to the number of candidates per landmark.

[5] Opposite to [Essafi *et al.* 2009a], we have considered a subset of 50 from the 895 model landmarks uniformly distributed in the model-space (Fig. 5.2).



(a) (b)

Figure 5.5: Experimental Settings for 3D Model Inference (a) The distribution of landmarks; (b) The histogram presenting the distribution of the number of missing 2D correspondences in the first experiment. Each bin represents the number of tests (vertical axis) that have the corresponding number of missing 2D correspondences (horizontal axis).

5.4.3 Results on 3D Model Inference from 2D images

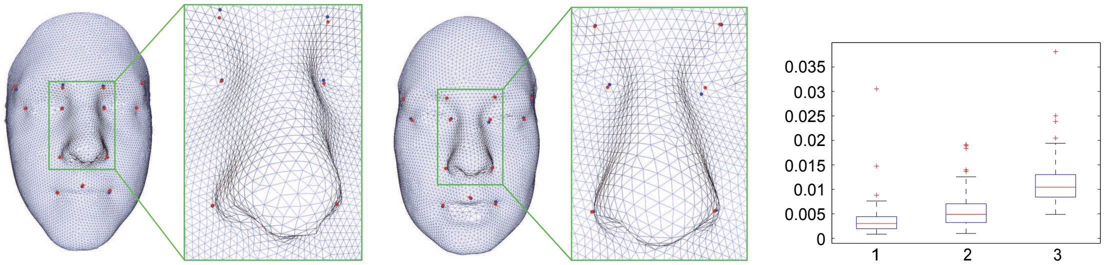
Experimental Setting

The performance of the proposed method was evaluated on the publicly-available facial expression datasets *BU-3DFE* [Yin *et al.* 2006] and *BU-4DFE* [Yin *et al.* 2008]. The former consists of 3D range data of 6 prototypical facial expressions of 100 different subjects (56 female and 44 male), and the latter is composed of 3D dynamic facial expressions of 101 different subjects (58 female and 43 male). The subjects included in both datasets are of various ethnic/racial origins.

The considered model consists of 13 landmarks (eyes, nose, mouth and eyebrows as shown in Fig. 5.5(a)). In the inference stage, its 3D initialization was done by randomly picking one training example. Regarding the 3D positions of the landmarks, the search was guided by a coarse-to-fine scheme and sparse sampling strategy in a similar way as [Glocker *et al.* 2008a]. Upon convergence of the algorithm, we performed *Procrustes Analysis* [Dryden & Mardia 1998] to obtain the similarity transform between the estimated 3D model and the ground truth, then transformed the estimated one into the referential frame of the ground truth. In terms of quantitative evaluation, a common goodness-of-fit criterion is the squared error standardized by the scale of the object. Thus, *Procrustes distance* [Dryden & Mardia 1998] was used as the dissimilarity measure E^{d} to evaluate our method quantitatively, which can be computed as follows:

$$E^{\mathrm{d}} = \sum_{i\in\mathcal{V}} \left\|\dot{x}_i^{(3)} - \hat{x}_i^{(3)}\right\|^2 \Big/ \sum_{i\in\mathcal{V}} \left\|\hat{x}_i^{(3)} - \hat{\mathbf{C}}^{(3)}\right\|^2 \tag{5.29}$$

where $\dot{x}_i^{(3)}$ and $\hat{x}_i^{(3)}$ denote the resulting and ground truth 3D positions of landmark i,



(a) (b) (c)

Figure 5.6: Results of the First Experiment. (a) and (b): 3D model estimation results. In each sub-figure, 3D face mesh is provided for measuring visually the error between the resulting positions (in red) of landmarks and the ground truth (in blue). (c): Boxplots for the distributions of dissimilarity measures for qualitatively evaluating the 3D model estimation. c.1: Results obtained by the proposed method; c.2: Results obtained by the version without visibility modeling; c.3: Initialization of the model. On each box, the central mark in red is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

respectively, $\hat{\mathbf{C}}^{(3)} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \hat{x}_i^{(3)}$ is the center of the ground truth model. The smaller E^{d} is, the closer the resulting model is to the ground truth.

In all the experiments, the concept of leave-one-out cross-validation was adopted towards the evaluation of the method. In this context, we do the validation on a sample while using the remaining samples as training data, and such a validation is done for all the samples contained in a dataset using the same parameter settings. Regarding the 3D model prior (Eq. 5.22), we modeled the probability distribution $p_{\mathbf{c}}(\hat{\mathbf{d}}_{\mathbf{c}}(\mathbf{x}_{\mathbf{c}}^{(3)}))$ between a quadruplet $\mathbf{c}$ of points using a two-component Gaussian Mixture.

Qualitative Results and Quantitative Analysis

First, we considered 100 samples of the neutral expression from *BU-3DFE*, one from each subject. The 2D landmark correspondence space was associated with 5 labels, four corresponding to the 2D position candidates and the last to the occlusion label "occ". On top of the ground truth correspondence, noise was added to generate erroneous 2D candidates as well. Furthermore, for 10% of the landmarks (randomly sampled), the true correspondence was removed and replaced with a random position in the image plane, which produced between 0 and 5 missing 2D correspondences for each test (see Fig. 5.5(b)). Figs. 5.6(a) and (b) present 3D model estimation results. Fig. 5.6(c).3 and Fig. 5.6(c).1 (*i.e.*, the boxes 3 and 1 in Fig. 5.6(c)) depict the statistics of the dissimilarity measure E^{d} (Eq. 5.29) for the initialization and the resulting 3D model obtained by the proposed method, respectively.

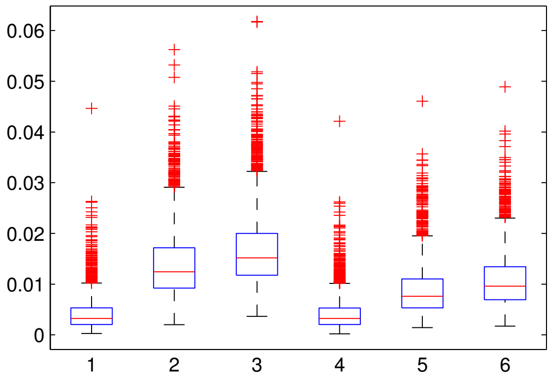


Figure 5.7: Comparison with ASM+RANSAC in term of Dissimilarity Measures. 1. Our method with random-sample initialization; 2. ASM+RANSAC with random-sample initialization; 3. Random-sample initialization; 4. Our method with mean-shape initialization; 5. ASM+RANSAC with mean-shape initialization; 6. Mean-shape initialization.

The qualitative and quantitative evaluations demonstrate that our method leads to well-estimated 3D models even when correspondences are partially missing. Furthermore, in order to demonstrate the impact of the visibility modeling, we have also evaluated an alternative version (without visibility modeling) of the proposed method where the “occ” label was removed from the 2D candidate set of each node, and show the obtained statistics of E^{d} in Fig. 5.6(c).2. Based on the comparison of Fig. 5.6(c).1 and Fig. 5.6(c).2, we can conclude that the visibility modeling indeed leads to significantly better performance.

Second, we employed the facial feature point detector of [Vukadinovic & Pantic 2005] to obtain the 2D position candidates for 101 samples of *BU-4DFE*, also one from each subject. Such a detector is based on Gabor features and boosting classifiers, and can well localize the considered landmarks from observed 2D images (Figs. 5.8(a)-(f)), though errors may still be present in some tests. We also performed a leave-one-out cross-validation in this experiment. Figs. 5.8(a')-(f') show six 3D model estimation results of different qualities and Fig. 5.8(g) presents the statistics of E^{d} for the proposed method and the version without visibility modeling. These results further demonstrate the potential of the proposed method to infer the 3D configuration of the model from 2D observed images with misdetections/occlusion handling.

Last but not least, we compared our method with an alternative method (ASM+RANSAC) with a relaxed condition where we assumed that the ground truth 2D correspondences were known. For each test, we first learned an ASM [Cootes *et al.* 1995] from the training data. Then, we used RANSAC [Fischler & Bolles 1981] to estimate the camera projection function based on the initialization of the shape model and the given ground truth 2D correspondences. Once the projection function was estimated, we searched for the best shape configuration by minimizing the errors between the projections of the 3D points and their 2D correspondences. Furthermore, we evaluated both methods using two different initializations: besides the "random sample" initialization used throughout the experiments, we also tested the "mean-shape" initialization where we chose one example as the reference, registered all the other training examples to it and computed the mean shape as initialization. We performed leave-one-out cross-validation on all the 2500 samples of *BU-3DFE* dataset and the quantitative evaluation is shown in Fig. 5.7. Figs. 5.7.1 and 5.7.4 show that our method performed equally well with the two different initializations, which demonstrates robustness with respect to the choice of initialization. The evaluation of ASM+RANSAC is presented in Figs. 5.7.2 and 5.7.5. We observe from Fig. 5.7 that the dissimilarity measure of our method is approximately 3 to 5 times lower compared to ASM+RANSAC, which demonstrates that our method performs significantly better than ASM+RANSAC and is highly robust with respect to the initialization.

In conclusion, the results of all the experiments demonstrate that our method, despite the important variability of pose and facial geometry, has well estimated the 3D configuration of the model even with the existence of misdetections, and outperforms significantly the alternative methods.

5.5 Conclusion

In this chapter, we have introduced one-shot optimization approaches for 3D knowledge-based segmentation and for 3D landmark model inference from a monocular 2D view based on higher-order MRFs, respectively. In order to eliminate potential effects of global pose estimation in the training and testing stages, the shape prior manifolds are built upon higher-order interactions of landmarks from a training set where pose-invariant statistics are obtained. In particular for the problem of 3D model inference from 2D images, the proposed 3D-2D consistency that is also encoded in such high-order interactions eliminate the necessity of viewpoint estimation, and the modeling of visibility improves further the performance of the method by handling missing correspondences and occlusions. The main innovations of the methods are the absence of global pose estimation and/or camera parameters estimation, the ability to model geometric consistency through local priors and the one-shot optimization to jointly infer all the variables. Furthermore, the explicit modeling of visibility in the 2D-3D inference formulation has been demonstrated to be able to handle missing correspondences and occlusion. Our methods have achieved promising results on challenging medical image data and standard facial datasets, respectively.

Incorporating regional and/or edge-based image support into the proposed MRF models will significantly enlarge their extent of applications. We have also studied the problem of decomposition of such image support in [Xiang *et al.* 2011, Wang *et al.* 2011a]. In knowledge-based 2D or 3D segmentation with MRFs, while edge-based terms are somewhat easy to be modeled in a distributed manner, it is not straightforward to decompose the regional data likelihood into local terms since such a regional term involves integrals on the regions which are delimited by the contour (depending on the positions of all the control points). In the work of [Xiang *et al.* 2011], which I have participated into, an exact factorization of the regional data term was proposed by using *divergence theorem* and leads to significantly better performance compared to the method of [Besbes *et al.* 2009] which relies on an approximative decomposition. The integration of such distributed regional terms in the proposed 3D pose-invariant segmentation framework will certainly yield powerful segmentation algorithms for many challenging scenarios such as 3D tagged magnetic resonance image segmentation, which is being under investigation. In [Wang *et al.* 2011a], we proposed an approach to deal with 3D reconstruction from bi-planar images given camera parameters, where regional and boundary likelihoods from 2D images are modeled using higher-order potentials. We are investigating an efficient and accurate approach to fuse similar distributed likelihood terms in the current joint 2D-3D inference framework.

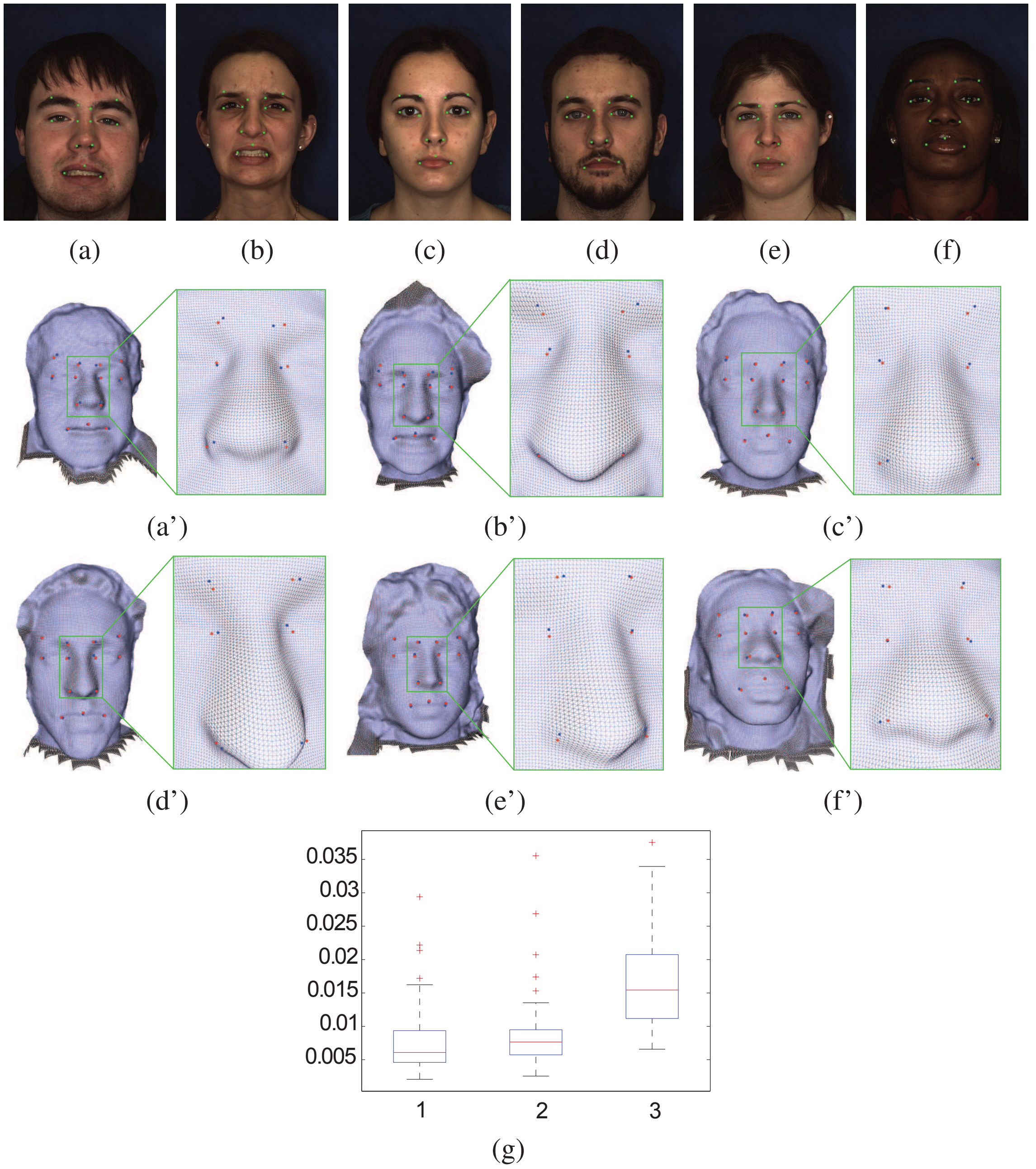


Figure 5.8: Results of the Second Experiment. (a)-(f): 2D landmark detection results [Vukadinovic & Pantic 2005]; (a')-(f'): The corresponding 3D model estimation results. (g): Boxplots for the distributions of dissimilarity measures for qualitatively evaluating the 3D model estimation. g.1: Results obtained by the proposed method; g.2: Results obtained by the version without visibility modeling; g.3: Initialization of the model.

Chapter 6

Conclusion

In this thesis, we have introduced graph-based modeling to address several fundamental problems of computer vision. In particular, our contributions refer to segmentation, tracking, shape matching and 3D model inference. The driving force of this thesis was the use of distributed and higher-order graphical models. Such a choice was motivated by the need of single-shot optimization methods that take into account the complementarity of visual perception tasks while at the same time inherit invariance with respect to certain global parameters such as the camera viewpoint and the object pose.

6.1 Contributions

The main contributions of this thesis are the following:

- We have proposed a joint 2.5D layered image representation. Opposite to classic 2.5D layered representations that require a high-order model due to the integration of depth ordering, we achieved a model that is restricted by local constraints involving only pairs of variables. Such a representation provides novel insights to solve multi-object motion estimation problems and allows to use a pairwise objective function to jointly model and solve depth ordering, segmentation and tracking.

  Then, based on this 2.5D layered representation, we have developed a single-shot optimization framework for joint segmentation, depth ordering and multi-object tracking using a pairwise MRF, where all the variables of interest interact. Furthermore, the fusion of depth ordering leads to a rigorous visibility modeling and occlusion handling for segmenting and tracking multiple objects. The proposed formulation is modular with respect to the data term and the shape representation, while being independent from the inference algorithm.

- Towards introducing richer geometric prior knowledge in the grouping problem, we have studied and proposed a novel algorithm for non-rigid 3D surface matching via a higher-order graph-based formulation that accounts for geometric/appearance similarities and intrinsic deformation errors. This was achieved through a third-order graph matching method where a pseudo-boolean objective function is optimized using a dual-decomposition-based approach together with recent order-reduction techniques, so as to achieve optimal correspondences between two surfaces. The principled fusion of the distortions at both local and global levels leads to a high robustness of the proposed method.

- We have introduced a pose-invariant distributed shape model whose prior manifold is described through accumulation of local densities involving pose invariant combinations of points. Then, based on such a statistical shape model, we have proposed a novel approach for knowledge-based 3D segmentation based on a higher-order MRF, where the prior information from the shape model and the image likelihood defined by classification techniques are combined together and the inference is done using a one-shot optimization through a dual-decomposition-based higher-order MRF optimization method.

- We have proposed a novel approach for 3D landmark model inference from a monocular 2D view that combines the estimation of the 3D model parameters, the visibility states and the 2D correspondences. The proposed probabilistic inference approach does not require explicit viewpoint estimation, while being able to jointly optimize the 3D model parameters and the corresponding landmarks selection as well as explicitly handling missing correspondences and occlusions via a visibility modeling. The image likelihood, the visibility prior, the 3D geometric prior and the 3D-2D projection consistency prior that compose the posterior probability are naturally modeled using a higher-order MRF and all the latent variables are inferred also through the dual-decomposition-based method.

The proposed formulations are modular with respect to the optimization method. We believe that their strength will become more and more significant with the development of optimization techniques. We also expect that our models could inspire the graph-based modeling for other vision problems.

6.2 Future Works

We now open the discussion on several directions of research that are related to the presented works and are or will be investigated.

Regarding the 2.5D layered model and the formulation for joint segmentation, tracking and depth-ordering, several important future directions are:

- As opposed to simple shape priors such as rectangles, we can imagine more complex object representations that are able to cope with important deformations (*e.g.*, point distribution models). One of the most promising directions for future work would be to incorporate the graph representation of shape models into the existing MRF model towards a more accurate understanding of the scene. For example, we are particularly interested in searching for a principled approach to combine our pose-invariant shape prior into the unified framework for joint segmentation, tracking and depth-ordering.

- Another promising direction is to extend the current framework to deal with articulated objects such as the pose estimation of human body and/or hand. A straightforward way is to model each component of articulated objects using an object node and add pairwise interactions between object nodes to model spatial constraints between the corresponding components (similar to *pictorial model*). The rigorous handling of visibility/occlusion of such a framework could greatly impact the quality of the obtained results.

- The extension of the joint 2.5D layered model to deal with the depth ordering problem in other related vision problems such as motion segmentation, layer decomposition and optical flow is also under investigation. Existing methods for these problems usually assume/impose that the layers are strictly and totally ordered according to their relative depths. Thus, the decomposition of the depth ordering into low-order interactions would lead to novel graph-based formulations to efficiently solve such problems.

- In each node of the proposed MRF model, the candidate set of the latent random variable is a product space. Theoretical questions are to be addressed in such a context like how to explore the structure of the energy function of such product-space MRF models and develop a more efficient optimization algorithm[1] both in terms of computational speed and memory requirement. Moreover, we believe that the development of such an efficient optimization algorithm would motivate new product-space MRF models for many vision problems.

[1] Note that [Goldluecke & Cremers 2010] have investigated the optimization of a class of product-space MRFs where the form of pairwise potentials are quite limited (*i.e.*, *separable metrics*).

Some future works related to the non-rigid 3D surface matching are as follows:

- A promising direction is to study shape similarity analysis, recognition and retrieval based on the surface matching method. The robust matching performance of our method could provide a strong cornerstone for these applications.

- The current matching formulation is based on the isometric assumption. The relaxation of this assumption towards handling wider deformation groups (*e.g.*, diffeomorphism) is an interesting direction. Moreover, the probabilistic extension of the matching formulation is also an important problem for dealing with the variability within a class of shapes.

- 3D surface tracking is also a promising direction which can be applied in various attractive applications such as facial expression analysis and transfer. We are particularly interested in developing a unified graph-based 3D surface tracking with advanced deformation priors.

Regarding the 3D model inference from 2D or 3D images, we are interested in the following directions:

- Better decomposition towards recovering the smallest subset of higher-order interactions that can express the 3D geometric manifold is a natural step forward. Such an approach could drastically decrease the computational complexity of the methods.

- Towards widening the application set of our knowledge-based segmentation and simultaneous 2D-3D estimation formulations, more advanced parameterizations of the manifold which go beyond simple 3D landmark positions (*e.g.*, the entire surface through some kind of local interpolation) could be employed. Besides, another promising direction is a principled incorporation of regional and/or edge-based image support in the current formulations, as we discussed in section 5.5.

- Faster optimization algorithms of higher-order MRFs could be beneficial to our approach both in terms of the considered application as well as in terms of modularity with respect to other 3D model inference problems. Hence, an important problem that needs to be dealt with in the future is the development of such a faster optimizer.

- An interesting future work is to develop a graph-based method to track 3D facial expression from monocular 2D images based on our techniques on the joint 2D-3D estimation and on the non-rigid 3D surface matching. We believe that a robust algorithm would be contributive to many applications related to facial expression.

- Last but not least, one promising direction is to extend the current formulations to the scenario of 3D model tracking. Towards this goal, an interesting problem is to model and incorporate spatio-temporal higher-order priors on the shape with dynamic behavior (*e.g.*, anatomical structure, face).

Publications of the Author

- Chaohui Wang, Martin de La Gorce and Nikos Paragios. *Segmentation, Ordering and Multi-object Tracking Using Graphical Models*. In IEEE International Conference on Computer Vision (ICCV), 2009.

- Chaohui Wang, Olivier Teboul, Fabrice Michel, Salma Essafi and Nikos Paragios. *3D Knowledge-Based Segmentation Using Pose-Invariant Higher-Order Graphs*. In International Conference, Medical Image Computing and Computer Assisted Intervention (MICCAI), 2010.

- Chaohui Wang, Yun Zeng, Loic Simon, Ioannis Kakadiaris, Dimitris Samaras and Nikos Paragios. *Viewpoint Invariant 3D Landmark Model Inference from Monocular 2D Images Using Higher-Order Priors*. In IEEE International Conference on Computer Vision (ICCV), 2011.

- Chaohui Wang, Haithem Boussaid, Loic Simon, Jean-Yves Lazennec and Nikos Paragios. *Pose-invariant 3D Proximal Femur Estimation through Bi-Planar Image Segmentation with Hierarchical Higher-Order Graph-based Priors*. In International Conference, Medical Image Computing and Computer Assisted Intervention (MICCAI), 2011.

- Chaohui Wang, Michael M. Bronstein, Alexander M. Bronstein and Nikos Paragios. *Discrete Minimum Distortion Correspondence Problems for Non-rigid Shape Matching*. In International Conference on Scale Space and Variational Methods in Computer Vision (SSVM), 2011. (**Oral presentation**)

- Yun Zeng, Chaohui Wang, Yang Wang, Xianfeng Gu, Dimitris Samaras and Nikos Paragios. *Dense non-rigid surface registration using high-order graph matching*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

- Yun Zeng, Chaohui Wang, Yang Wang, Xianfeng Gu, Dimitris Samaras and Nikos Paragios. *Intrinsic Dense 3D Surface Tracking*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. (**Oral presentation**)

- Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras and Nikos Paragios. *Illumination Estimation and Cast Shadow Detection through a Higher-order Graphical Model*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

- Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras and Nikos Paragios. *Estimating Shadows with the Bright Channel Cue*. In Color and Reflectance in Imaging and Computer Vision Workshop (CRICV) (in conjuction with ECCV), 2010.

- Bo Xiang, Chaohui Wang, Jean-Francois Deux, Alain Rahmouni and Nikos Paragios. *Tagged Cardiac MR Image Segmentation Using Boundary & Regional-Support and Graph-based Deformable Priors*. In IEEE International Symposium on Biomedical Imaging (ISBI), 2011. (**Oral presentation**)

Bibliography

[Agarwal & Triggs 2004] Ankur Agarwal and Bill Triggs. *3D Human Pose from Silhouettes by Relevance Vector Regression*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004. 56

[Aji & McEliece 2000] S.M. Aji and R.J. McEliece. *The generalized distributive law*. IEEE Transactions on Information Theory, vol. 46, no. 2, pp. 325–343, March 2000. 35, 40, 43, 44

[Alahari *et al.* 2008] Karteek Alahari, Pushmeet Kohli and Philip H. S. Torr. *Reduce, reuse & recycle: Efficiently solving multi-label MRFs*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 39

[Ali *et al.* 2008] Asem M. Ali, Aly A. Farag and Georgy L. Gimel'farb. *Optimizing binary MRFs with higher order cliques*. In European Conference on Computer Vision (ECCV), 2008. 50

[Alter 1994] T. D. Alter. *3-D pose from 3 points using weak-perspective*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 16, no. 8, pp. 802–808, 1994. 114, 115, 116

[Amini *et al.* 1990] Amir A. Amini, Terry E. Weymouth and Ramesh C. Jain. *Using dynamic programming for solving variational problems in vision*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 12, no. 9, pp. 855–867, 1990. 31

[Amit & Geman 1997] Yali Amit and Donald Geman. *Shape quantization and recognition with randomized trees*. Neural computation, vol. 9, no. 7, pp. 1545–1588, 1997. 109

[Andriluka *et al.* 2009] Mykhaylo Andriluka, Stefan Roth and Bernt Schiele. *Pictorial structures revisited: People detection and articulated pose estimation*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 30

[Anguelov *et al.* 2004] Dragomir Anguelov, Praveen Srinivasan, Hoi-Cheung Pang, Daphne Koller, Sebastian Thrun and James Davis. *The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces*. In Advances in Neural Information Processing Systems (NIPS), 2004. 79

[Aurenhammer 1991] Franz Aurenhammer. *Voronoi diagrams - a survey of a fundamental geometric data structure*. ACM Computing Surveys, vol. 23, no. 3, pp. 345–405, 1991. 103

[Auvray *et al.* 2009] Vincent Auvray, Patrick Bouthemy and Jean Liénard. *Joint Motion Estimation and Layer Segmentation in Transparent Image Sequences - Application to Noise Reduction in X-Ray Image Sequences*. EURASIP Journal on Advances in Signal Processing, vol. 2009, pp. 19:1–19:21, 2009. 56

[Balan & Black 2006] Alexandru O. Balan and Michael J. Black. *An Adaptive Appearance Model Approach for Model-based Articulated Object Tracking*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. 54

[Balan *et al.* 2007] Alexandru O. Balan, Leonid Sigal, Michael J. Black, James E. Davis and Horst W. Haussecker. *Detailed Human Shape and Pose from Images*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007. 105

[Batra *et al.* 2010] Dhruv Batra, A. C. Gallagher, Devi Parikh and Tsuhan Chen. *Beyond Trees: MRF Inference via Outer-Planar Decomposition*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 45, 48

[Bellman 1957] R. Bellman. Dynamic Programming. Princeton University Press, 1957. 30, 40, 102

[Belongie *et al.* 2002] S. Belongie, J. Malik and J. Puzicha. *Shape matching and object recognition using shape contexts*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 24, no. 4, pp. 509–522, 2002. 81

[Berg *et al.* 2005] Alexander C. Berg, Tamara L. Berg and Jitendra Malik. *Shape Matching and Object Recognition Using Low Distortion Correspondences*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005. 81

[Bernardino & Santos-Victor 2006] Alexandre Bernardino and José Santos-Victor. *Fast IIR Isotropic 2-D Complex Gabor Filters With Boundary Initialization*. IEEE Transactions on Image Processing (TIP), vol. 15, no. 11, pp. 3338–3348, 2006. 110

[Bertsekas 1999] Dimitri P. Bertsekas. Nonlinear Programming (Second Edition). Athena Scientific, 1999. 47, 51, 83, 88, 105, 119

[Besag 1974] Julian Besag. *Spatial Interaction and the Statistical Analysis of Lattice Systems*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 36, no. 2, pp. 192–236, 1974. 25

[Besag 1986] Julian Besag. *On the Statistical Analysis of Dirty Pictures Julian Besag (with discussion)*. Journal of the Royal Statistical Society (Series B), vol. 48, no. 3, pp. 259–302, 1986. 35

[Besbes *et al.* 2009] A. Besbes, N. Komodakis, G. Langs and N. Paragios. *Shape priors and discrete MRFs for knowledge-based segmentation*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 54, 102, 103, 126

[Besl & McKay 1992] Paul J. Besl and Neil D. McKay. *A Method for Registration of 3-D Shapes*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 14, no. 2, pp. 239–256, 1992. 78

[Bishop 1995] Christopher M. Bishop. Neural networks for pattern recognition. Oxford university press, 1995. 24

[Bishop 2006] Christopher M. Bishop. Pattern recognition and machine learning (Information Science and Statistics). Springer, 2006. 23, 30, 34, 40, 105

[Black & Anandan 1993] Michael J. Black and P. Anandan. *A framework for the robust estimation of optical flow*. In IEEE International Conference on Computer Vision (ICCV), 1993. 28

[Blake & Zisserman 1987] Andrew Blake and Andrew Zisserman. Visual Reconstruction. MIT Press, 1987. 35

[Blanz & Vetter 1999] Volker Blanz and Thomas Vetter. *A morphable model for the synthesis of 3D faces*. In SIGGRAPH, pp. 187–194, 1999. 105, 106

[Blemker *et al.* 2007] Silvia S. Blemker, Deanna S. Asakawa, Garry E. Gold and Scott L. Delp. *Image-based musculoskeletal modeling: Applications, advances, and future opportunities*. Journal of magnetic resonance imaging, vol. 25, no. 2, pp. 441–451, 2007. 104

[Blum 1973] Harry Blum. *Biological shape and visual science*. Journal of Theoretical Biology, vol. 38, no. 2, pp. 205–287, 1973. 101

[Bookstein 1989] Fred L. Bookstein. *Principal warps: Thin-plate splines and the decomposition of deformations*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 11, no. 6, pp. 567–585, 1989. 104, 121

[Borgefors 1986] Gunilla Borgefors. *Distance transformations in digital images*. Computer vision, graphics, and image processing, vol. 34, no. 3, pp. 344–371, 1986. 51

[Boros & Hammer 2002] Endre Boros and Peter L. Hammer. *Pseudo-boolean optimization*. Discrete Applied Mathematics, vol. 123, pp. 155–225, 2002. 37, 50, 83, 85, 88

[Boros *et al.* 1991] Endre Boros, P. L. Hammer and X. Sun. *Network flows and minimization of quadratic pseudo-Boolean functions*. RUTCOR Research Report RRR 17-1991, 1991. 35, 39, 85

[Boros *et al.* 2006] Endre Boros, P. L. Hammer and Gabriel Tavares. *Preprocessing of unconstrained quadratic binary optimization*. RUTCOR Research Report RRR 10-2006, 2006. 35, 39, 85

[Bosch *et al.* 2007] Anna Bosch, Andrew Zisserman and Xavier Munoz. *Image classification using random forests and ferns*. IEEE International Conference on Computer Vision (ICCV), 2007. 109

[Boser *et al.* 1992] Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik. *An training algorithm for optimal margin classifiers*. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152, 1992. 66, 109

[Boyd & Vandenberghe 2004] Stephen Boyd and Lieven Vandenberghe. Convex Optimization. Cambridge University Press, 2004. 46

[Boykov & Funka-Lea 2006] Yuri Boykov and Gareth Funka-Lea. *Graph Cuts and Efficient N-D Image Segmentation*. International Journal of Computer Vision (IJCV), vol. 70, no. 2, pp. 109–131, November 2006. 27, 28, 36, 54, 65, 104

[Boykov & Jolly 2001] Yuri Boykov and Marie-Pierre Jolly. *Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images*. In IEEE International Conference on Computer Vision (ICCV), 2001. 33

[Boykov & Kolmogorov 2003] Yuri Boykov and Vladimir Kolmogorov. *Computing geodesics and minimal surfaces via graph cuts*. In IEEE International Conference on Computer Vision (ICCV), 2003. 28, 35

[Boykov *et al.* 1998] Yuri Boykov, Olga Veksler and Ramin Zabih. *Markov random fields with efficient approximations*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1998. 29, 31, 35, 38

[Boykov *et al.* 2001] Yuri Boykov, Olga Veksler and Ramin Zabih. *Fast Approximate Energy Minimization via Graph Cuts*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 23, no. 11, pp. 1222–1239, 2001. 27, 28, 35, 38, 55

[Bregler *et al.* 2000] Christoph Bregler, Aaron Hertzmann and Henning Biermann. *Recovering non-rigid 3D shape from image streams*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2000. 105

[Breiman 1996] Leo Breiman. *Bagging predictors*. Machine Learning, vol. 24, no. 2, pp. 123–140, 1996. 109

[Breiman 2001] Leo Breiman. *Random Forests*. Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. 66, 105, 109, 114

[Brejl & Sonka 2000] Marek Brejl and Milan Sonka. *Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples*. IEEE Transactions on Medical Imaging (TMI), vol. 19, no. 10, pp. 973–985, 2000. 104

[Bronstein & Kokkinos 2010] M. M. Bronstein and I. Kokkinos. *Scale-invariant heat kernel signatures for non-rigid shape recognition*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 86

[Bronstein *et al.* 2006] A. M. Bronstein, M. M. Bronstein and R. Kimmel. *Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching*. Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 103, no. 5, pp. 1168–1172, 2006. 79, 85

[Bronstein *et al.* 2007] Alexander M. Bronstein, Michael M. Bronstein and Ron Kimmel. *Expression-Invariant Representations of Faces*. IEEE Transactions on Image Processing (TIP), vol. 16, no. 1, pp. 188–197, 2007. 79

[Bronstein *et al.* 2010] A. M. Bronstein, M. M. Bronstein, R. Kimmel, M. Mahmoudi and G. Sapiro. *A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching*. International Journal of Computer Vision (IJCV), vol. 89, no. 2–3, pp. 266–286, 2010. 79, 85

[Brown & Rusinkiewicz 2007] Benedict J. Brown and Szymon Rusinkiewicz. *Global non-rigid alignment of 3-D scans*. ACM Transactions on Graphics (TOG), vol. 26, no. 3, 2007. 78

[Buatois *et al.* 2009] Luc Buatois, Guillaume Caumon and Bruno Levy. *Concurrent number cruncher: a GPU implementation of a general sparse linear solver*. International Journal of Parallel, Emergent and Distributed Systems, vol. 24, no. 3, pp. 205–223, 2009. 94

[Caelli & Kosinov 2004] T. Caelli and S. Kosinov. *An eigenspace projection clustering method for inexact graph matching*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 26, no. 4, pp. 515–519, 2004. 82

[Campbell & Flynn 2001] Richard J. Campbell and Patrick J. Flynn. *A survey of free-form object representation and recognition techniques*. Computer Vision and Image Understanding (CVIU), vol. 81, no. 2, 2001. 77

[Carcassoni & Hancock 2003] M. Carcassoni and E. R. Hancock. *Spectral correspondence for point pattern matching*. Pattern Recognition, vol. 36, no. 1, pp. 193–204, January 2003. 82

[Carr & Hartley 2009] Peter Carr and Richard Hartley. *Minimizing energy functions on 4-connected lattices using elimination*. In IEEE International Conference on Computer Vision (ICCV), 2009. 39

[Caselles *et al.* 1997] Vicent Caselles, Ron Kimmel and Guillermo Sapiro. *Geodesic active contours*. International Journal of Computer Vision (IJCV), vol. 22, no. 1, pp. 61–79, January 1997. 54

[Chan & Shen 2005] Tony F. Chan and Jianhong Shen. Image processing and analysis: variational, PDE, wavelet, and stochastic methods. Society for Industrial and Applied Mathematics (SIAM), 2005. 29

[Chan & Vese 2001] Tony F. Chan and Luminita A. Vese. *Active contours without edges*. IEEE Transactions on Image Processing (TIP), vol. 10, no. 2, pp. 266–277, 2001. 54

[Chen *et al.* 2010] Yu Chen, Tae-Kyun Kim and Roberto Cipolla. *Inferring 3D shapes and deformations from single views*. In European Conference on Computer Vision (ECCV), 2010. 102, 105

[Chertok & Keller 2010] Michael Chertok and Yosi Keller. *Efficient high order matching*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 32, no. 12, pp. 2205–2215, December 2010. 82

[Chou & Brown 1990] Paul B. Chou and Christopher M. Brown. *The theory and practice of Bayesian image labeling*. International Journal of Computer Vision (IJCV), vol. 4, no. 3, pp. 185–210, 1990. 35

[Chou *et al.* 1993] P. B. Chou, P. R. Cooper, M. J. Swain, C. M. Brown and L. E. Wixson. *Probabilistic network inference for cooperative high and low level vision*. In R. Chellappa and A. Jain, editeurs, Markov Random Fields: Theory and Applications, pp. 211–243, 1993. 35

[Comaniciu & Meer 2002] Dorin Comaniciu and Peter Meer. *Mean shift: A robust approach toward feature space analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 24, no. 5, pp. 603–619, 2002. 54, 91

[Comaniciu *et al.* 2000] D. Comaniciu, V. Ramesh and P. Meer. *Real-time tracking of non-rigid objects using mean shift*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2000. 54, 55

[Comaniciu *et al.* 2003] Dorin Comaniciu, Visvanathan Ramesh and Peter Meer. *Kernel-Based Object Tracking*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 25, no. 5, pp. 564–575, 2003. 54

[Conte *et al.* 2004] D. Conte, P. Foggia, C. Sansone and M. Vento. *Thirty years of graph matching in pattern recognition*. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), vol. 18, no. 3, pp. 265–298, 2004. 81

[Cootes & Taylor 1999] T. F. Cootes and C. J. Taylor. *A mixture model for representing shape variation*. Image and Vision Computing (IVC), vol. 17, no. 8, pp. 567–573, June 1999. 102

[Cootes *et al.* 1995] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham. *Active Shape Models-Their Training and Application*. Computer Vision and Image Understanding (CVIU), vol. 61, no. 1, pp. 38–59, January 1995. 54, 101, 124

[Cootes *et al.* 2001] T. F. Cootes, G. J. Edwards and C. J. Taylor. *Active appearance models*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 23, no. 6, pp. 681–685, 2001. 101

[Cordella *et al.* 1996] L. P. Cordella, P. Foggia, C. Sansone and M. Vento. *An efficient algorithm for the inexact matching of arg graphs using a contextual transformational model*. In International Conference on Pattern Recognition (ICPR), 1996. 82

[Cordella *et al.* 2001] L. P. Cordella, P. Foggia, C. Sansone and M. Vento. *An improved algorithm for matching large graphs*. In 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, 2001. 82

[Cormen *et al.* 2009] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest and Clifford Stein. Introduction to algorithms (Third Edition). The MIT press, 2009. 26, 30, 40, 44, 102

[Cortes & Vapnik 1995] Corinna Cortes and Vladimir N. Vapnik. *Support-vector networks*. Machine Learning, vol. 20, no. 3, pp. 273–297, 1995. 66, 109

[Cour *et al.* 2007] Timothee Cour, Praveen Srinivasan and Jianbo Shi. *Balanced graph matching*. In Advances in Neural Information Processing Systems (NIPS), 2007. 82

[Cremers *et al.* 2007] Daniel Cremers, Mikael Rousson and Rachid Deriche. *A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape*. International Journal of Computer Vision (IJCV), vol. 72, no. 2, pp. 195–215, August 2007. 101

[Dahlhaus *et al.* 1992] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour and M. Yannakakis. *The complexity of multiway cuts (extended abstract)*. In ACM Symposium on Theory of Computing (STOC), 1992. 36

[Darrell & Fleet 1995] Trevor Darrell and David Fleet. *Second-order method for occlusion relationships in motion layers*. Technical Report 314, MIT Media Lab, 1995. 56, 59

[Dawid 1992] A. P. Dawid. *Applications of a general propagation algorithm for probabilistic expert systems*. Statistics and Computing, vol. 2, no. 1, pp. 25–36, 1992. 35, 40, 43

[de Berg *et al.* 2000] Mark de Berg, M. van Krefeld, M. Overmars and O. Schwarzkopf. Computational geometry: Algorithms and applications. Springer, 2 édition, 2000. 92

[de La Gorce *et al.* 2008] Martin de La Gorce, Nikos Paragios and David J. Fleet. *Model-based hand tracking with texture, shading and self-occlusions*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 54

[de La Gorce *et al.* 2011] Martin de La Gorce, David J. Fleet and Nikos Paragios. *Model-based 3d hand pose estimation from monocular video*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 33, no. 9, pp. 1793–1805, 2011. 101, 105

[Delong *et al.* 2010] Andrew Delong, Anton Osokin, Hossam N. Isack and Yuri Boykov. *Fast approximate energy minimization with label costs*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 32, 49

[Dinic & Kronrod 1969] E. A. Dinic and M. A. Kronrod. *An Algorithm for the Solution of the Assignment Problem*. Soviet Mathematics Doklady, vol. 10, pp. 1324–1326, 1969. 81

[do Carmo 1976] Manfredo P. do Carmo. Differential geometry of curves and surfaces. Prentice Hall, 1976. 85, 87

[Donner *et al.* 2007] Rene Donner, Branislav Micusík, Georg Langs and Horst Bischof. *Sparse MRF appearance models for fast anatomical structure localisation*. In British Machine Vision Conference (BMVC), 2007. 104

[Dryden & Mardia 1998] I. L. Dryden and K. V. Mardia. Statistical shape analysis. John Wiley & Sons Inc., 1998. 101, 122

[Dubrovina & Kimmel 2010] A. Dubrovina and R. Kimmel. *Matching shapes by eigendecomposition of the Laplace-Beltrami operator*. In International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), 2010. 86

[Duchenne *et al.* 2009] O. Duchenne, F. Bach, I. Kweon and J. Ponce. *A tensor-based algorithm for high-order graph matching*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 82, 89

[Egmont-Petersen *et al.* 2002] M. Egmont-Petersen, D. de Ridderb and H. Handelsc. *Image processing with neural networks-a review*. Pattern Recognition, vol. 35, no. 10, pp. 2279–2301, October 2002. 24

[Eichner & Ferrari 2009] Marcin Eichner and Vittorio Ferrari. *Better appearance models for pictorial structures*. In British Machine Vision Conference (BMVC), 2009. 30

[Elad & Kimmel 2001] A. Elad and R. Kimmel. *Bending invariant representations for surfaces*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 168–174, 2001. 79

[Engl *et al.* 1996] H. W. Engl, M. Hanke and A. Neubauer. Regularization of inverse problems. Kluwer Academic Publishers, Dordrecht, 1996. 22, 29

[Essafi *et al.* 2009a] Salma Essafi, G. Langs, J-F. Deux, A. Rahmouni, G. Bassez and N. Paragios. *Wavelet-driven knowledge-based MRI calf muscle segmentation*. IEEE International Symposium on Biomedical Imaging (ISBI), 2009. 103, 104, 120, 121

[Essafi *et al.* 2009b] Salma Essafi, Georg Langs and Nikos Paragios. *Hierarchical 3D diffusion wavelet shape priors*. In IEEE International Conference on Computer Vision (ICCV), 2009. 101

[Farkas & Kra 2004] Hershel M. Farkas and Irwin Kra. Riemann surfaces. Springer, 2004. 86

[Felzenszwalb & Huttenlocher 2005] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. *Pictorial Structures for Object Recognition*. International Journal of Computer Vision (IJCV), vol. 61, no. 1, pp. 55–79, January 2005. 28, 30, 54

[Felzenszwalb & Huttenlocher 2006] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. *Efficient Belief Propagation for Early Vision*. International Journal of Computer Vision (IJCV), vol. 70, no. 1, pp. 41–54, May 2006. 35, 40, 41, 51

[Felzenszwalb & Mcauley 2011] Pedro F. Felzenszwalb and Julian J. Mcauley. *Fast Inference with Min-Sum Matrix Product*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 33, no. 12, pp. 2549–2554, 2011. 51

[Felzenszwalb & Zabih 2011] Pedro F. Felzenszwalb and Ramin Zabih. *Dynamic programming and graph algorithms in computer vision*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 33, no. 4, pp. 721–740, April 2011. 40

[Felzenszwalb *et al.* 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 (TPAMI), vol. 32, no. 9, pp. 1627–1645, September 2010. 30

[Fischler & Bolles 1981] Martin A. Fischler and Robert C. Bolles. *Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography*. Communications of the ACM, vol. 24, no. 6, pp. 381–395, 1981. 114, 124

[Fischler & Elschlager 1973] M.A. Fischler and R.A. Elschlager. *The representation and matching of pictorial structures*. IEEE Transactions on Computers, vol. 22, no. 1, pp. 67–92, 1973. 30

[Ford & Fulkerson 1962] L. R. Ford and D. R. Fulkerson. Flows in networks. Princeton University Press, 1962. 37

[Forsyth *et al.* 2005] David A. Forsyth, Okan Arikan, Leslie Ikemoto, James O'Brien and Deva Ramanan. *Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis*. Foundations and Trends in Computer Graphics and Vision, vol. 1, no. 2-3, pp. 77–254, 2005. 105

[Freedman & Drineas 2005] Daniel Freedman and Petros Drineas. *Energy Minimization via Graph Cuts: Settling What is Possible*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005. 50, 85

[Freedman & Zhang 2005] D. Freedman and T. Zhang. *Interactive Graph Cut Based Segmentation with Shape Priors*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005. 54, 55

[Freeman *et al.* 2000] William T. Freeman, Egon C. Pasztor and Owen T. Carmichael. *Learning low-level vision*. International Journal of Computer Vision (IJCV), vol. 40, no. 1, pp. 25–47, 2000. 35, 40, 41

[Freund & Schapire 1997] Y. Freund and R. Schapire. *A desicion-theoretic generalization of on-line learning and an application to boosting*. Journal of Computer and System Sciences (JCSS), vol. 55, no. 1, pp. 119–139, 1997. 66, 109

[Frey & MacKay 1998] Brendan J. Frey and David J. C. MacKay. *A revolution: Belief propagation in graphs with cycles*. In Advances in Neural Information Processing Systems (NIPS), 1998. 40

[Frey 1998] Brendan J. Frey. Graphical models for machine learning and digital communication. MIT Press, 1998. 34, 105

[Gallagher *et al.* 2011] Andrew C. Gallagher, Dhruv Batra and Devi Parikh. *Inference for Order Reduction in Markov Random Fields*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. 50

[Geiger *et al.* 1990] Dan Geiger, Thomas Verma and Judea Pearl. *Identifying independence in bayesian networks*. Networks, vol. 20, pp. 507–534, 1990. 24

[Gelb 1974] Arthur Gelb, editeur. Applied optimal estimation. MIT Press, 1974. 24, 55

[Geman & Geman 1984] Stuart Geman and Donald Geman. *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 6, no. 6, pp. 721–741, 1984. 28, 29, 33, 35

[Globerson & Jaakkola 2007] Amir Globerson and Tommi Jaakkola. *Fixing Max-Product: Convergent Message Passing Algorithms for MAP LP-Relaxations*. In Advances in Neural Information Processing Systems (NIPS), 2007. 46

[Glocker *et al.* 2008a] Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab and Nikos Paragios. *Dense image registration through MRFs and efficient linear programming*. Medical Image Analysis, vol. 12, no. 6, pp. 731–741, December 2008. 27, 28, 68, 122

[Glocker *et al.* 2008b] Ben Glocker, Nikos Paragios, Nikos Komodakis, Georgios Tziritas and Nassir Navab. *Optical flow estimation with uncertainties through dynamic MRFs*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 22

[Glocker *et al.* 2010] Ben Glocker, T. Hauke Heibel, Nassir Navab, Pushmeet Kohli and Carsten Rother. *TriangleFlow: Optical Flow with Triangulation-Based Higher-Order Likelihoods*. In European Conference on Computer Vision (ECCV), 2010. 31

[Gold & Rangarajan 1996] S. Gold and A. Rangarajan. *A graduated assignment algorithm for graph matching*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 18, no. 4, pp. 377–388, April 1996. 82

[Goldberg & Tarjan 1988] A. V. Goldberg and R. E. Tarjan. *A new approach to the maximum-flow problem*. Journal of the ACM (JACM), vol. 35, no. 4, pp. 921–940, 1988. 37

[Goldluecke & Cremers 2010] Bastian Goldluecke and Daniel Cremers. *Convex relaxation for multilabel problems with product label spaces*. In European Conference on Computer Vision (ECCV), 2010. 131

[Grady & Schwartz 2006] Leo Grady and Eric L. Schwartz. *Isoperimetric graph partitioning for image segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 28, no. 3, pp. 469–475, 2006. 54

[Greig *et al.* 1989] D. M. Greig, B. T. Porteous and A. H. Seheult. *Exact Maximum A Posteriori Estimation for Binary Images*. Journal of the Royal Statistical Society (Series B), vol. 51, no. 2, pp. 271–279, 1989. 28, 35, 36, 38

[Gromov 1981] M. Gromov. Structures Métriques Pour les Variétés Riemanniennes. Textes Mathématiques. 1981. 79

[Gu & Kanade 2006] Leon Gu and Takeo Kanade. *3d alignment of face in a single image*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. 105, 106

[Gu & Kanade 2008] Leon Gu and Takeo Kanade. *A generative shape regularization model for robust face alignment*. In European Conference on Computer Vision (ECCV), 2008. 102

[Guan *et al.* 2009] Peng Guan, Alexander Weiss, Alexandru O. Balan and Michael J. Black. *Estimating Human Shape and Pose from a Single Image*. In IEEE International Conference on Computer Vision (ICCV), 2009. 105

[Gui *et al.* 2008] Laura Gui, Jean-Philippe Thiran and Nikos Paragios. *Cooperative Object Segmentation and Behavior Inference in Image Sequences*. International Journal of Computer Vision (IJCV), vol. 84, no. 2, pp. 146–162, June 2008. 24

[Hahnel *et al.* 2003] Dirk Hahnel, Sebastian Thrun and Wolfram Burgard. *An extension of the ICP algorithm for modeling nonrigid objects with mobile robots*. In International Joint Conference on Artificial Intelligence (IJCAI), pp. 915–920, 2003. 78

[Hall 1935] Philip Hall. *On Representatives of Subsets*. Journal of the London Mathematical Society, vol. 10, pp. 26–30, 1935. 81

[Hammer *et al.* 1984] P. L. Hammer, P. Hansen and B. Simeone. *Roof duality, complementation and persistency in quadratic 0-1 optimization*. Mathematical Programming, vol. 28, no. 2, pp. 121–155, 1984. 39

[Hammersley & Clifford 1971] J. M. Hammersley and P. Clifford. *Markov fields on finite graphs and lattices*. unpublished, 1971. 25

[He *et al.* 2004] Xuming He, Richard S. Zemel and Miguel A. Carreira-Perpinan. *Multiscale conditional random fields for image labeling*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004. 33

[Hernández *et al.* 2007] Carlos Hernández, George Vogiatzis, Gabriel J. Brostow, Bjorn Stenger and Roberto Cipolla. *Non-rigid photometric stereo with colored lights*. In IEEE International Conference on Computer Vision (ICCV), 2007. 77

[Hervieu *et al.* 2007] A. Hervieu, P. Bouthemy and J.-P. Le Cadre. *A HMM-based method for recognizing dynamic video contents from trajectories*. In IEEE International Conference on Image Processing (ICIP), 2007. 24

[Ho 1998] Tin Kam Ho. *The random subspace method for constructing decision forests*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 20, no. 8, pp. 832–844, 1998. 109

[Hopcroft & Karp 1973] John E. Hopcroft and Richard M. Karp. *An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs*. SIAM Journal on Computing (SICOMP), vol. 2, pp. 225–231, 1973. 81

[Hu & Hua 2009] J. Hu and J. Hua. *Salient spectral geometric features for shape matching and retrieval*. Visual Computer, vol. 25, no. 5, pp. 667–675, 2009. 86

[Huang & Essa 2005] Y. Huang and I. Essa. *Tracking Multiple Objects through Occlusions*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005. 55, 56, 70

[Huang *et al.* 2004] R. Huang, V. Pavlovic and D. N. Metaxas. *A Graphical Model Framework for Coupling MRFs and Deformable Models*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004. 54, 55

[Huang *et al.* 2008] Qi-Xing Huang, Bart Adams, Martin Wicke and Leonidas J. Guibas. *Non-rigid registration under isometric deformations*. In Proceedings of the Eurographics Symposium on Geometry Processing (SGP), pp. 1449–1457, 2008. 78

[Huang *et al.* 2011] Yuchi Huang, Qingshan Liu and Dimitris N Metaxas. *A component-based framework for generalized face alignment*. IEEE Transactions on Systems, Man and Cybernetics (TSMC), vol. 41, no. 1, pp. 287–298, March 2011. 102

[Iannizzotto & Vita 2000] Giancarlo Iannizzotto and Lorenzo Vita. *Fast and accurate edge-based segmentation with no contour smoothing in 2-D real images*. IEEE Transactions Image Processing (TIP), vol. 9, no. 7, pp. 1232–1237, 2000. 104

[Isard & Blake 1998] Michael Isard and Andrew Blake. *CONDENSATION - conditional density propagation for visual tracking*. International Journal of Computer Vision (IJCV), vol. 29, no. 1, pp. 5–28, August 1998. 55

[Isard 2003] Michael Isard. *PAMPAS: real-valued graphical models for computer vision*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2003. 21, 30

[Ishikawa & Geiger 1998] Hiroshi Ishikawa and Davi Geiger. *Segmentation by grouping junctions*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1998. 35, 36, 38

[Ishikawa 2003] Hiroshi Ishikawa. *Exact optimization for Markov random fields with convex priors*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 25, no. 10, pp. 1333–1336, 2003. 27, 37, 38, 39

[Ishikawa 2009] Hiroshi Ishikawa. *Higher-order clique reduction in binary graph cut*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2009. 49, 50, 82, 83, 85, 88

[Ishikawa 2011] Hiroshi Ishikawa. *Transformation of General Binary MRF Minimization to the First Order Case*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 33, no. 6, pp. 1234–1249, March 2011. 50

[Ising 1925] E. Ising. *Beitrag zur theorie des ferromagnetismus*. Zeitschrift fur Physik, vol. 31, no. 1, pp. 253–258, 1925. 29

[Jackson *et al.* 2008] Jeremy D. Jackson, Anthony J. Yezzi and Stefano Soatto. *Dynamic Shape and Appearance Modeling via Moving and Deforming Layers*. International Journal of Computer Vision (IJCV), vol. 79, no. 1, pp. 71–84, December 2008. 56

[Jepson *et al.* 2002] Allan D. Jepson, David J. Fleet and Michael J. Black. *A Layered Motion Representation with Occlusion and Compact Spatial Support*. In European Conference on Computer Vision (ECCV), 2002. 56, 59

[Johnson 1997] Andrew Johnson. *Spin-Images: A Representation for 3-D Surface Matching*. PhD thesis, CMU, 1997. 86

[Jojic & Frey 2001] Nebojsa Jojic and Brendan J. Frey. *Learning Flexible Sprites in Video Layers*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2001. 56

[Jojic *et al.* 2010] Vladimir Jojic, Stephen Gould and Daphne Koller. *Accelerated dual decomposition for MAP inference*. In International Conference on Machine Learning (ICML), 2010. 49

[Jolliffe 2002] I.T. Jolliffe. Principal Component Analysis. Springer, 2002. 101

[Jordan 2007] Michael I. Jordan. An introduction to probabilistic graphical models. In preparation, 2007. 23, 25, 35, 40, 43, 44, 45

[Juan & Boykov 2006] Olivier Juan and Yuri Boykov. *Active Graph Cuts*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. 38

[Kalman 1960] R. E. Kalman. *A New Approach to Linear Filtering and Prediction Problems*. Transactions of the ASME - Journal of Basic Engineering, vol. 82, pp. 35–45, 1960. 24, 55

[Kappes *et al.* 2010] Jörg Hendrik Kappes, Stefan Schmidt and Christoph Schnörr. *MRF Inference by k-Fan Decomposition and Tight Lagrangian Relaxation*. In European Conference on Computer Vision (ECCV), 2010. 48

[Kass *et al.* 1988] Michael Kass, Andrew Witkin and Demetri Terzopoulos. *Snakes: Active contour models*. International Journal of Computer Vision (IJCV), vol. 1, no. 4, pp. 321–331, January 1988. 31, 54, 104

[Kim & Woods 1997] Jaemin Kim and J. W. Woods. *Spatio-temporal adaptive 3-D Kalman filter for video*. IEEE Transactions on Image Processing (TIP), vol. 6, no. 3, pp. 414–424, January 1997. 24

[Kinect 2010] Kinect. *Microsoft© Kinect*, 2010. 77

[Kjæ rulff 1998] Uffe Kjæ rulff. *Inference in bayesian networks using nested junction trees*. In Proceedings of the NATO Advanced Study Institute on Learning in graphical models, 1998. 45

[Kohli & Pawan Kumar 2010] Pushmeet Kohli and M. Pawan Kumar. *Energy minimization for linear envelope MRFs*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 52

[Kohli & Torr 2005] Pushmeet Kohli and Philip H. S. Torr. *Efficiently solving dynamic Markov random fields using graph cuts*. IEEE International Conference on Computer Vision (ICCV), 2005. 38, 55

[Kohli & Torr 2007] Pushmeet Kohli and Philip H. S. Torr. *Dynamic graph cuts for efficient inference in Markov Random Fields*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 29, no. 12, pp. 2079–2088, December 2007. 27, 28, 38

[Kohli & Torr 2008] Pushmeet Kohli and Philip H. S. Torr. *Measuring uncertainty in graph cut solutions*. Computer Vision and Image Understanding (CVIU), vol. 112, no. 1, pp. 30–38, October 2008. 22

[Kohli *et al.* 2007] Pushmeet Kohli, M. Pawan Kumar and Philip H. S. Torr. *P3 & Beyond: Solving Energies with Higher Order Cliques*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007. 31, 49, 50

[Kohli *et al.* 2008a] Pushmeet Kohli, Lubor Ladicky and Philip H. S. Torr. *Robust higher order potentials for enforcing label consistency*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 31, 50

[Kohli *et al.* 2008b] Pushmeet Kohli, Jonathan Rihan, Matthieu Bray and Philip H. S. Torr. *Simultaneous Segmentation and Pose Estimation of Humans Using Dynamic Graph Cuts*. International Journal of Computer Vision (IJCV), vol. 79, no. 3, pp. 285–298, January 2008. 27, 36, 54, 55, 65, 104

[Kohli *et al.* 2008c] Pushmeet Kohli, Alexander Shekhovtsov, Carsten Rother, Vladimir Kolmogorov and Philip H. S. Torr. *On partial optimality in multi-label MRFs*. In International Conference on Machine Learning (ICML), 2008. 39

[Kohli *et al.* 2009a] Pushmeet Kohli, Lubor Ladicky and Philip H. S. Torr. *Robust Higher Order Potentials for Enforcing Label Consistency*. International Journal of Computer Vision (IJCV), vol. 82, no. 3, pp. 302–324, January 2009. 31, 50

[Kohli *et al.* 2009b] Pushmeet Kohli, M. Pawan Kumar and Philip H. S. Torr. *P3 & beyond: move making algorithms for solving higher order functions*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 31, no. 9, pp. 1645–1656, September 2009. 31, 50, 51, 52

[Kokkinos & Yuille 2008] I. Kokkinos and A. Yuille. *Scale Invariance without Scale Selection*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 110

[Koller & Friedman 2009] Daphne Koller and Nir Friedman. Probabilistic graphical models: Principles and techniques. The MIT Press, 2009. 23, 24

[Kolmogorov & Rother 2007] Vladimir Kolmogorov and Carsten Rother. *Minimizing nonsubmodular functions with graph cuts - a review*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 29, no. 7, pp. 1274–1279, July 2007. 35, 39, 85, 88

[Kolmogorov & Wainwright 2005] Vladimir Kolmogorov and Martin J. Wainwright. *On the optimality of tree-reweighted max-product message-passing*. In Conference on Uncertainty in Artificial Intelligence (UAI), 2005. 47, 48

[Kolmogorov & Zabih 2002] Vladimir Kolmogorov and Ramin Zabih. *Multi-camera Scene Reconstruction via Graph Cuts*. In European Conference on Computer Vision (ECCV), 2002. 27, 28, 35, 36

[Kolmogorov & Zabih 2004] Vladimir Kolmogorov and Ramin Zabih. *What energy functions can be minimized via graph cuts?* IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 26, no. 2, pp. 147–159, February 2004. 27, 35, 36, 37, 38, 50

[Kolmogorov 2006] Vladimir Kolmogorov. *Convergent tree-reweighted message passing for energy minimization*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 28, no. 10, pp. 1568–1583, October 2006. 22, 27, 28, 35, 36, 40, 41, 46, 47, 69, 120

[Komodakis & Paragios 2008] Nikos Komodakis and Nikos Paragios. *Beyond Loose LP-relaxations: Optimizing MRFs by Repairing Cycles*. In European Conference on Computer Vision (ECCV), 2008. 46

[Komodakis & Paragios 2009] Nikos Komodakis and Nikos Paragios. *Beyond pairwise energies: Efficient optimization for higher-order MRFs*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 31, 49, 51, 119, 120

[Komodakis & Tziritas 2007] Nikos Komodakis and Georgios Tziritas. *Approximate labeling via graph cuts based on linear programming*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 29, no. 8, pp. 1436–1453, August 2007. 38

[Komodakis *et al.* 2007a] Nikos Komodakis, Nikos Paragios and Georgios Tziritas. *MRF Optimization via Dual Decomposition: Message-Passing Revisited*. In IEEE International Conference on Computer Vision (ICCV), 2007. 22, 28, 41, 47, 83, 88, 105, 119, 120

[Komodakis *et al.* 2007b] Nikos Komodakis, Georgios Tziritas and Nikos Paragios. *Fast, Approximately Optimal Solutions for Single and Dynamic MRFs*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007. 35, 36, 38, 46, 51

[Komodakis *et al.* 2008] Nikos Komodakis, Georgios Tziritas and Nikos Paragios. *Performance vs computational efficiency for optimizing single and dynamic MRFs: Setting the state of the art with primal-dual strategies*. Computer Vision and Image Understanding (CVIU), vol. 112, no. 1, pp. 14–29, October 2008. 27, 35, 38

[Komodakis *et al.* 2011] Nikos Komodakis, Nikos Paragios and Georgios Tziritas. *MRF energy minimization and beyond via dual decomposition*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 33, no. 3, pp. 531–552, March 2011. 36, 46, 47, 48

[Komodakis 2010] Nikos Komodakis. *Towards More Efficient and Effective LP-Based Algorithms for MRF Optimization*. In European Conference on Computer Vision (ECCV), 2010. 27, 49

[Koval & Schlesinger 1976] V. K. Koval and M. I. Schlesinger. *Dvumernoe programmirovanie v zadachakh analiza izobrazheniy (Two-dimensional programming in image analysis problems)*. USSR Academy of Science, Automatics and Telemechanics, vol. 8, pp. 149–168, 1976. 46

[Kovalevsky & Koval 1975] V. A. Kovalevsky and V. K. Koval. *A diffusion algorithm for decreasing energy of max-sum labeling problem*. Technical report, Glushkov Institute Of Cybernetics, Kiev, USSR, 1975. 46, 51

[Kschischang *et al.* 2001] Frank R. Kschischang, Brendan J. Frey and Hans-Andrea Loeliger. *Factor graphs and the sum-product algorithm*. IEEE Transactions on Information Theory, vol. 47, no. 2, pp. 498–519, 2001. 34, 40

[Kuhn 1955] Harold W. Kuhn. *The Hungarian Method for the assignment problem*. Naval Research Logistics Quarterly, vol. 2, pp. 83–97, 1955. 81

[Kumar & Hebert 2004] Sanjiv Kumar and Martial Hebert. *Discriminative fields for modeling spatial dependencies in natural images*. In Advances in Neural Information Processing Systems (NIPS), 2004. 33

[Kurazume *et al.* 2009] Ryo Kurazume, Kaori Nakamura, Toshiyuki Okada, Yoshinobu Sato, Nobuhiko Sugano, Tsuyoshi Koyama, Yumi Iwashita and Tsutomu

Hasegawa. *3D reconstruction of a femoral shape using a parametric model and two 2D fluoroscopic images*. Computer Vision and Image Understanding (CVIU), vol. 113, no. 2, pp. 202–211, 2009. 105

[Kwon *et al.* 2008] Dongjin Kwon, Kyong Joon Lee, Il Dong Yun and Sang Uk Lee. *Non-rigid Image Registration Using Dynamic Higher-Order MRF Model*. In European Conference on Computer Vision (ECCV), 2008. 31

[Ladicky *et al.* 2009] Lubor Ladicky, Christopher Russell, Pushmeet Kohli and Philip H. S. Torr. *Associative hierarchical CRFs for object class image segmentation*. In IEEE International Conference on Computer Vision (ICCV), 2009. 33

[Ladicky *et al.* 2010a] Lubor Ladicky, Christopher Russell, Pushmeet Kohli and Philip H. S. Torr. *Graph Cut based Inference with Co-occurrence Statistics*. In European Conference on Computer Vision (ECCV), 2010. 32, 49, 50

[Ladicky *et al.* 2010b] Lubor Ladicky, Paul Sturgess, Karteek Alahari, Chris Russell and Philip H. S. Torr. *What, where & how many? combining object detectors and crfs*. European Conference on Computer Vision (ECCV), 2010. 33

[Ladicky *et al.* 2011] Lubor Ladicky, Christopher Russell, Pushmeet Kohli and Philip H. S. Torr. *Inference Methods for CRFs with Co-occurrence Statistics*. International Journal of Computer Vision (IJCV), 2011. 32, 50

[Lafferty *et al.* 2001] John D. Lafferty, Andrew McCallum and Fernando C. N. Pereira. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In International Conference on Machine Learning (ICML), 2001. 32

[Lan *et al.* 2006] Xiangyang Lan, Stefan Roth, Daniel P. Huttenlocher and Michael J. Black. *Efficient Belief Propagation with Learned Higher-Order Markov Random Fields*. In European Conference on Computer Vision (ECCV), 2006. 49, 51

[Lauritzen 1996] S. L. Lauritzen. Graphical Models. Oxford University Press, 1996. 23

[Lawrence 2004] Neil D. Lawrence. *Gaussian process latent variable models for visualisation of high dimensional data*. In Advances in Neural Information Processing Systems (NIPS), 2004. 102

[Lee & Liu 1999] R.S.T. Lee and J.N.K. Liu. *An oscillatory elastic graph matching model for recognition of offline handwritten Chinese characters*. In International Conference Knowledge-Based Intelligent Information Engineering Systems, 1999. 81

[Lee & Pavlidis 1988] D. Lee and T. Pavlidis. *One-dimensional regularization with discontinuities*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 10, no. 6, pp. 822–829, 1988. 29

[Lempitsky *et al.* 2010] Victor Lempitsky, Carsten Rother, Stefan Roth and Andrew Blake. *Fusion moves for markov random field optimization*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 32, no. 8, pp. 1392–1405, August 2010. 39, 50

[Leordeanu & Hebert 2005] Marius Leordeanu and Martial Hebert. *A spectral technique for correspondence problems using pairwise constraints*. In IEEE International Conference on Computer Vision (ICCV), 2005. 81, 82

[Leotta & Mundy 2011] M. J. Leotta and J. L. Mundy. *Vehicle surveillance with a generic, adaptive, 3-D vehicle model*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 33, no. 7, pp. 1457–1469, 2011. 105

[Lepetit & Fua 2006] Vincent Lepetit and Pascal Fua. *Keypoint recognition using randomized trees.* IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 28, no. 9, pp. 1465–1479, September 2006. 109

[Li 2009] Stan Z. Li. Markov random field modeling in image analysis (Third Edition). Springer, 2009. 26

[Lipman & Funkhouser 2009] Yaron Lipman and Thomas Funkhouser. *Möbius voting for surface correspondence*. ACM Transactions on Graphics (TOG), vol. 28, no. 3, pp. 72:1–72:12, 2009. 80, 83, 86, 89, 90, 94

[Lovasz & Plummer 1986] L. Lovasz and M. D. Plummer. Matching Theory. North-Holland Mathematics Studies, 1986. 81

[Malcolm *et al.* 2007] J. Malcolm, Y. Rathi and A. Tannenbaum. *Multi-Object Tracking Through Clutter Using Graph Cuts*. In IEEE International Conference on Computer Vision (ICCV), 2007. 55

[Markelj *et al.* 2010] P. Markelj, D. Tomazevic, B. Likar and F. Pernus. *A review of 3D/2D registration methods for image-guided interventions.* Medical Image Analysis, April 2010. 105

[Mateus *et al.* 2008] D. Mateus, R. P. Horaud, D. Knossow, F. Cuzzolin and E. Boyer. *Articulated shape matching using Laplacian eigenfunctions and unsupervised point registration*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 86

[Mcauley & Caetano 2011] Julian J. Mcauley and Tiberio S. Caetano. *Faster Algorithms for Max-Product Message-Passing*. Journal of Machine Learning Research, vol. 12, pp. 1349–1388, 2011. 51

[Mémoli & Sapiro 2005] F. Mémoli and G. Sapiro. *A theoretical and computational framework for isometry invariant recognition of point cloud data*. Foundations of Computational Mathematics, vol. 5, no. 3, pp. 313–347, 2005. 79, 85

[Moni & Ali 2009] M. A. Moni and A. B. M. Shawkat Ali. *HMM based hand gesture recognition: A review on techniques and approaches*. In IEEE International Conference on Computer Science and Information Technology (ICCSIT), 2009. 24

[Moore *et al.* 2010] Alastair P. Moore, Simon J. D. Prince and Jonathan Warrell. *"Lattice Cut" - Constructing superpixels using layer constraints*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 28

[Mueller *et al.* 2003] K. Mueller, A. Smolic, M. Droese, P. Voigt and T. Wienand. *Multitexture modeling of 3D traffic scenes*. In International Conference on Multimedia and Expo (ICME), 2003. 105

[Muller *et al.* 2001] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda and B. Scholkopf. *An introduction to kernel-based learning algorithms*. IEEE Transactions on Neural Networks, vol. 12, no. 2, pp. 181–201, 2001. 66, 109

[Mumford & Shah 1989] D. Mumford and J. Shah. *Optimal approximations by piecewise smooth functions and associated variational problems*. Communications on Pure and Applied Mathematics, vol. 42, no. 5, pp. 577–685, 1989. 54

[Nitzberg & Mumford 1990] M. Nitzberg and D. Mumford. *The 2.1-D sketch*. In IEEE International Conference on Computer Vision (ICCV), 1990. 56

[Nowozin & Lampert 2009] Sebastian Nowozin and Christoph H. Lampert. *Global connectivity potentials for random field models*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 32, 49

[O'Rourke & Badler 1980] J. O'Rourke and N. I. Badler. *Model-based image analysis of human motion using constraint propagation*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 2, no. 6, pp. 522–536, 1980. 105

[Osher & Fedkiw 2002] Stanley Osher and Ronald P. Fedkiw. Level set methods and dynamic implicit surfaces. Springer, 2002. 54, 67, 101

[Osher & Paragios 2003] Stanley Osher and Nikos Paragios, editeurs. Geometric level set methods in imaging, vision, and graphics. Springer, 2003. 54

[Ovsjanikov *et al.* 2009] M. Ovsjanikov, A. M. Bronstein, M. M. Bronstein and L. J. Guibas. *Shape Google: a computer vision approach to invariant shape retrieval*. In Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment (NORDIA), 2009. 86

[Panagopoulos *et al.* 2010] Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras and Nikos Paragios. *Estimating Shadows with the Bright Channel Cue*. In Color and Reflectance in Imaging and Computer Vision Workshop (CRICV) (in conjuction with ECCV), 2010. 74

[Panagopoulos *et al.* 2011] Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras and Nikos Paragios. *Illumination Estimation and Cast Shadow Detection through a Higher-order Graphical Model*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. 74, 75

[Paragios & Deriche 2002] Nikos Paragios and Rachid Deriche. *Geodesic active regions and level set methods for supervised texture segmentation*. International Journal of Computer Vision (IJCV), vol. 46, no. 3, pp. 223–247, 2002. 54, 104

[Paragios *et al.* 2005] Nikos Paragios, Yunmei Chen and Olivier Faugeras. Handbook of Mathematical Models in Computer Vision. Springer-Verlag New York, Inc., 2005. 16

[Paskin 2003] Mark A. Paskin. *Thin Junction Tree Filters for Simultaneous Localization and Mapping*. In International Joint Conference on Artificial Intelligence (IJCAI), 2003. 45

[Pavlovic 1999] Vladimir Pavlovic. *Dynamic Bayesian Networks for Information Fusion with Application to Human-Computer Interfaces*. PhD thesis, University of Illinois at Urbana-Champaign, 1999. 24

[Pawan Kumar & Torr 2006] M. Pawan Kumar and Philip H. S. Torr. *Fast memory-efficient generalized belief propagation*. European Conference on Computer Vision (ECCV), 2006. 41

[Pawan Kumar *et al.* 2004] M. Pawan Kumar, Philip H. S. Torr and Andrew Zisserman. *Learning Layered Pictorial Structures from Video*. In The Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), 2004. 30

[Pawan Kumar *et al.* 2005] M. Pawan Kumar, Philip H. S. Torr and Andrew Zisserman. *OBJ CUT*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005. 55

[Pawan Kumar *et al.* 2008] M. Pawan Kumar, Philip H. S. Torr and Andrew Zisserman. *Learning Layered Motion Segmentations of Video*. International Journal of Computer Vision (IJCV), vol. 76, no. 3, pp. 301–319, 2008. 56

[Pawan Kumar *et al.* 2009] M. Pawan Kumar, Vladimir Kolmogorov and Philip H.S. Torr. *An Analysis of Convex Relaxations for MAP Estimation of Discrete MRFs*. Journal of Machine Learning Research, vol. 10, pp. 71–106, 2009. 27, 46

[Pawan Kumar 2008] M. Pawan Kumar. *Combinatorial and Convex Optimization for Probabilistic Models in Computer Vision*. PhD thesis, Oxford Brookes University, 2008. 32

[Pearl 1988] Judea Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, 1988. 24, 30, 34, 35, 40

[Petersen *et al.* 2008] Kersten Petersen, Janis Fehr and Hans Burkhardt. *Fast generalized belief propagation for MAP estimation on 2D and 3D grid-like markov random fields*. DAGM-Symposium, pp. 41–50, 2008. 41

[Pighin *et al.* 1998] Frederic Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski and David H. Salesin. *Synthesizing realistic facial expressions from photographs*. In SIGGRAPH, 1998. 105

[Pinkall & Polthier 1993] Ulrich Pinkall and Konrad Polthier. *Computing Discrete Minimal Surfaces and Their Conjugates*. Experimental Mathematics, vol. 2, no. 1, pp. 15–36, 1993. 94

[Pizer *et al.* 2003] Stephen M. Pizer, P. Thomas Fletcher, Sarang Joshi, Andrew Thall, James Z. Chen, Yonatan Fridman, Daniel S. Fritsch, A. Graham Gash, John M. Glotzer, Michael R. Jiroutek, Conglin Lu, Keith E. Muller, Gregg Tracton, Paul Yushkevich and Edward L. Chaney. *Deformable m-reps for 3d medical image segmentation*. International Journal of Computer Vision (IJCV), vol. 55, no. 2-3, pp. 85–106, 2003. 101

[Potetz & Lee 2008] Brian Potetz and Tai Sing Lee. *Efficient belief propagation for higher-order cliques using linear constraint nodes*. Computer Vision and Image Understanding (CVIU), vol. 112, no. 1, pp. 39–54, October 2008. 49, 51

[Potetz 2007] Brian Potetz. *Efficient Belief Propagation for Vision Using Linear Constraint Nodes*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007. 51

[Potts 1952] R. B. Potts. *Some generalized order-disorder transitions*. Proceedings of the Cambridge Philosophical Society, vol. 48, pp. 106–109, 1952. 29, 74

[Press *et al.* 1988] W. Press, B. Flannery, S. Teukolsky and W. Vetterling. Numerical recipes in c. Cambridge University Press, 1988. 55

[Qiu & Hancock 2007] H. Qiu and E. R. Hancock. *Clustering and Embedding Using Commute Times*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 29, no. 11, pp. 1873–1890, 2007. 85

[Quattoni *et al.* 2004] Ariadna Quattoni, Michael Collins and Trevor Darrell. *Conditional random fields for object recognition*. In Advances in Neural Information Processing Systems (NIPS), 2004. 33

[Rabiner 1989] L.R. Rabiner. *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989. 24

[Ramalingam *et al.* 2008] Srikumar Ramalingam, Pushmeet Kohli, Karteek Alahari and Philip H. S. Torr. *Exact inference in multi-label CRFs with higher order cliques*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 49

[Rocha & Pavlidis 1994] J. Rocha and T. Pavlidis. *A Shape Analysis Model with Applications to a Character Recognition System*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 16, pp. 393–404, April 1994. 81

[Roller *et al.* 1993] D. Roller, Kostas Daniilidis and Hans-Hellmut Nagel. *Model-based object tracking in monocular image sequences of road traffic scenes*. International Journal of Computer Vision (IJCV), vol. 10, no. 3, pp. 257–281, 1993. 105

[Romberg *et al.* 2001] J K Romberg, H Choi and R G Baraniuk. *Bayesian tree-structured image modeling using wavelet-domain hidden Markov models*. IEEE Transactions on Image Processing (TIP), vol. 10, no. 7, pp. 1056–1068, January 2001. 24

[Romdhani *et al.* 1999] Sami Romdhani, Shaogang Gong and Alexandra Psarrou. *A multi-view nonlinear active shape model using kernel PCA*. In British Machine Vision Conference (BMVC), 1999. 102

[Rosenberg 1975] I. G. Rosenberg. *Reduction of bivalent maximization to the quadratic case*. Cahiers du Centre d'etudes de Recherche Operationnelle, vol. 17, pp. 71–74, 1975. 50

[Roth & Black 2005] Stefan Roth and Michael J. Black. *Fields of Experts: A Framework for Learning Image Priors*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005. 31, 49

[Roth & Black 2007] Stefan Roth and Michael J. Black. *On the Spatial Statistics of Optical Flow*. International Journal of Computer Vision (IJCV), vol. 74, no. 1, pp. 33–50, January 2007. 22

[Roth & Black 2009] Stefan Roth and Michael J. Black. *Fields of Experts*. International Journal of Computer Vision (IJCV), vol. 82, no. 2, pp. 205–229, January 2009. 31, 49

[Rother *et al.* 2004] Carsten Rother, Vladimir Kolmogorov and Andrew Blake. *GrabCut - Interactive Foreground Extraction using Iterated Graph Cuts*. ACM Transactions on Graphics (TOG), vol. 23, no. 3, pp. 309–314, 2004. 28, 35, 36, 54, 104

[Rother *et al.* 2007] Carsten Rother, Vladimir Kolmogorov, Victor Lempitsky and Martin Szummer. *Optimizing Binary MRFs via Extended Roof Duality*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2007. 35, 39, 85

[Rother *et al.* 2009] Carsten Rother, Pushmeet Kohli, Wei Feng and Jiaya Jia. *Minimizing sparse higher order energy functions of discrete variables*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 52

[Roy & Cox 1998] Sebastien Roy and Ingemar J. Cox. *A Maximum-Flow Formulation of the N -camera Stereo Correspondence Problem*. In IEEE International Conference on Computer Vision (ICCV), 1998. 28, 35, 38

[Rusinkiewicz & Levoy 2001] Szymon Rusinkiewicz and Marc Levoy. *Efficient variants of the ICP algorithm*. In International Conference on 3-D Digital Imaging and Modeling, 2001. 78

[Rustamov 2007] R. M. Rustamov. *Laplace-Beltrami eigenfunctions for deformation invariant shape representation*. In Proceedings of the Eurographics Symposium on Geometry Processing (SGP), pp. 225–233, 2007. 86

[Sahni & Gonzalez 1976] S. Sahni and T. Gonzalez. *P-complete approximation problems*. Journal of the ACM (JACM), vol. 23, no. 3, pp. 555–565, 1976. 81

[Salakhutdinov 2009] Ruslan Salakhutdinov. *Learning in Markov random fields using tempered transitions*. In Advances in Neural Information Processing Systems (NIPS), 2009. 22

[Salzmann & Fua 2010] Mathieu Salzmann and Pascal Fua. *Linear Local Models for Monocular Reconstruction of Deformable Surfaces*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 33, no. 5, pp. 931–944, 2010. 105

[Sandhu *et al.* 2009] Romeil Sandhu, Samuel Dambreville, Anthony Yezzi and Allen Tannenbaum. *Non-rigid 2D-3D pose estimation and 2D image segmentation*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 106

[Schapire 1990] Robert E. Schapire. *The strength of weak learnability*. Machine Learning, vol. 5, no. 2, pp. 197–227, 1990. 66, 109

[Schapire 2001] Robert E. Schapire. *The boosting approach to machine learning: An overview*. In MSRI Workshop on Nonlinear Estimation and Classification, 2001. 66, 109

[Schellewald & Schnorr 2005] Christian Schellewald and Christoph Schnorr. *Probabilistic subgraph matching based on convex relaxation*. In Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), 2005. 82

[Schlesinger & Flach 2006] Dmitrij Schlesinger and Boris Flach. *Transforming an arbitrary minsum problem into a binary one*. Technical report, Dresden University of Technology, 2006. 35, 37

[Scholkopf *et al.* 1998] Bernhard Scholkopf, Alexander Smola and Klaus-Robert Muller. *Nonlinear component analysis as a kernel eigenvalue problem*. Neural Computation, vol. 10, no. 5, pp. 1299–1319, 1998. 102

[Schroff *et al.* 2008] F Schroff, A Criminisi and A Zisserman. *Object class segmentation using random forests*. In British Machine Vision Conference (BMVC), 2008. 109

[Seghers *et al.* 2007a] Dieter Seghers, Dirk Loeckx, Frederik Maes, Dirk Vandermeulen and Paul Suetens. *Minimal shape and intensity cost path segmentation*. IEEE Transactions on Medical Imaging (TMI), vol. 26, no. 8, pp. 1115–1129, August 2007. 102, 104

[Seghers *et al.* 2007b] Dieter Seghers, Pieter Slagmolen, Yves Lambelin, Jeroen Hermans, Frederik Maes and Paul Suetens. *Landmark based liver segmentation using* *local shape and local intensity models*. In 3D segmentation in the clinic: a grand challenge, 2007. 102, 104

[Senior *et al.* 2006] Andrew W. Senior, Arun Hampapur, Ying li Tian, Lisa M. G. Brown, Sharath Pankanti and Ruud M. Bolle. *Appearance models for occlusion handling*. Image and Vision Computing (IVC), vol. 24, no. 11, pp. 1233–1243, 2006. 56

[Shachter 1998] Ross D. Shachter. *Bayes-Ball: The Rational Pastime (for Determining Irrelevance and Requisite Information in Belief Networks and Influence Diagrams)*. In Conference on Uncertainty in Artificial Intelligence (UAI), 1998. 24

[Shaji *et al.* 2010] A. Shaji, A. Varol, L. Torresani and P. Fua. *Simultaneous Point Matching and 3D Deformable Surface Reconstruction*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 77

[Sheffer *et al.* 2006] Alla Sheffer, Emil Praun and Kenneth Rose. *Mesh parameterization methods and their applications*. Foundations and Trends in Computer Graphics and Vision, vol. 2, no. 2, pp. 105–171, 2006. 91, 95

[Shekhovtsov *et al.* 2008] Alexander Shekhovtsov, Ivan Kovtun and Vaclav Hlavac. *Efficient MRF deformation model for non-rigid image matching*. Computer Vision and Image Understanding (CVIU), vol. 112, no. 1, pp. 91–99, October 2008. 28

[Shi & Malik 2000] Jianbo Shi and Jitendra Malik. *Normalized cuts and image segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 22, no. 8, pp. 888–905, 2000. 54

[Sigal & Black 2006a] Leonid Sigal and Michael J. Black. *Measure Locally, Reason Globally: Occlusion-sensitive Articulated Pose Estimation*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. 30, 54, 56, 101

[Sigal & Black 2006b] Leonid Sigal and Michael J. Black. *Predicting 3D People from 2D Pictures*. In International Conference on Articulated Motion and Deformable Objects (AMDO), 2006. 105

[Sigal *et al.* 2003] Leonid Sigal, Michael Isard, Benjamin H. Sigelman and Michael J. Black. *Attractive People: Assembling Loose-Limbed Models using Non-parametric Belief Propagation*. In Advances in Neural Information Processing Systems (NIPS), 2003. 21, 30

[Sigal *et al.* 2007] Leonid Sigal, Alexandru O. Balan and Michael J. Black. *Combined discriminative and generative articulated pose and non-rigid shape estimation*. In Advances in Neural Information Processing Systems (NIPS), 2007. 105

[Simon *et al.* 2011] Loic Simon, Olivier Teboul, Panagiotis Koutsourakis and Nikos Paragios. *Random Exploration of the Procedural Space for Single-View 3D Modeling of Buildings*. International Journal of Computer Vision (IJCV), vol. 93, no. 2, pp. 253–271, 2011. 105, 109

[Smith *et al.* 2004] Paul Smith, Tom Drummond and Roberto Cipolla. *Layered Motion Segmentation and Depth Ordering by Tracking Edges*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 26, no. 4, pp. 479–494, April 2004. 56, 59

[Sontag & Jaakkola 2007] David Sontag and Tommi Jaakkola. *New outer bounds on the marginal polytope*. In Advances in Neural Information Processing Systems (NIPS), 2007. 46

[Staib & Duncan 1996] Lawrence H. Staib and James S. Duncan. *Model-based deformable surface finding for medical images*. IEEE Transactions on Medical Imaging (TMI), vol. 15, no. 5, pp. 720–731, 1996. 101

[Starner *et al.* 1998] Thad Starner, Joshua Weaver and Alex Pentland. *A wearable computer based american sign language recognizer*. In Assistive Technology and Artificial Intelligence: Applications in Robotics, User Interfaces and Natural Language Processing, pp. 84–96. Springer-Verlag, 1998. 24

[Stauffer & Grimson 1999] Chris Stauffer and W.E.L. Grimson. *Adaptive Background Mixture Models for Real-Time Tracking*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1999. 68

[Strandmark & Kahl 2010] Petter Strandmark and Fredrik Kahl. *Parallel and distributed graph cuts by dual decomposition*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 49

[Sudderth *et al.* 2004a] Erik B. Sudderth, Michael I. Mandel, William T. Freeman and Alan S. Willsky. *Distributed occlusion reasoning for tracking with nonparametric belief propagation*. In Advances in Neural Information Processing Systems (NIPS), 2004. 30, 54, 56, 112

[Sudderth *et al.* 2004b] Erik B. Sudderth, Michael I. Mandel, William T. Freeman and Alan S. Willsky. *Visual Hand Tracking Using Nonparametric Belief Propagation*. IEEE CVPR Workshop on Generative Model Based Vision, 2004. 30

[Sudderth *et al.* 2010] Erik B. Sudderth, Alexander T. Ihler, Michael Isard, William T. Freeman and Alan S. Willsky. *Nonparametric belief propagation*. Communications of the ACM, vol. 53, no. 10, pp. 95–103, 2010. 21, 30

[Sumner & Popović 2004] Robert W. Sumner and Jovan Popović. *Deformation transfer for triangle meshes*. ACM Transactions on Graphics (TOG), vol. 23, no. 3, pp. 399–405, 2004. 94

[Sun *et al.* 2003] Jian Sun, Nan-ning Zheng and Heung-yeung Shum. *Stereo matching using belief propagation*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 25, no. 7, pp. 787–800, July 2003. 35, 41

[Sun *et al.* 2009] J. Sun, M. Ovsjanikov and L. J. Guibas. *A concise and provably informative multi-scale signature based on heat diffusion*. In Computer Graphics Forum, volume 28, pp. 1383–1392, 2009. 86

[Sun *et al.* 2010] Deqing Sun, Erik B. Sudderth and Michael J. Black. *Layered Image Motion with Explicit Occlusions , Temporal Consistency , and Depth Ordering*. In Advances in Neural Information Processing Systems (NIPS), 2010. 28, 56

[Sutton & McCallum 2011] Charles Sutton and Andrew McCallum. *An Introduction to Conditional Random Fields*. Foundations and Trends in Machine Learning (To appear), 2011. 32

[Szeliski *et al.* 2008] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen and Carsten Rother. *A comparative study of energy minimization methods for Markov random fields with smoothness-based priors*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 30, no. 6, pp. 1068–1080, June 2008. 27, 41

[Szeliski 2010] Richard Szeliski. Computer vision: algorithms and applications. Springer-Verlag New York Inc., 2010. 15, 17, 22

[Tao *et al.* 2000] Hai Tao, Harpreet S. Sawhney and Rakesh Kumar. *Dynamic Layer Representation with Applications to Tracking*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2000. 56

[Tappen & Freeman 2003] Marshall F. Tappen and William T. Freeman. *Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters*. IEEE International Conference on Computer Vision (ICCV), 2003. 41

[Tarlow *et al.* 2010] Daniel Tarlow, Inmar E. Givoni and Richard S. Emel. *HOP-MAP: Efficient Message Passing with High Order Potentials*. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2010. 51

[Terzopoulos & Szeliski 1993] Demetri Terzopoulos and Richard Szeliski. *Tracking with Kalman snakes*. In Active vision, pp. 3–20. MIT Press, 1993. 24

[Terzopoulos 1986] Demetri Terzopoulos. *Regularization of inverse visual problems involving discontinuities*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 8, no. 4, pp. 413–424, 1986. 29

[Tevs *et al.* 2009] Art Tevs, Martin Bokeloh, Michael Wand, Andreas Schilling and Hans-Peter Seidel. *Isometric Registration of Ambiguous and Partial Data*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 94

[Thorstensen & Keriven 2009] N. Thorstensen and R. Keriven. *Non-rigid shape matching using Geometry and Photometry*. In Asian Conference on Computer Vision (ACCV), 2009. 85, 86

[Tikhonov & Arsenin 1977] A. N. Tikhonov and V. Y. Arsenin. Solutions of ill-posed problems. Winston Washington, DC:, 1977. 22, 29

[Torr 2003] Philip H. S. Torr. *Solving Markov Random Fields using Semi Definite Programming*. In Ninth International Workshop on Artificial Intelligence and Statistics, 2003. 82

[Torresani *et al.* 2008] Lorenzo Torresani, Vladimir Kolmogorov and Carsten Rother. *Feature Correspondence via Graph Matching: Models and Global Optimization*. In European Conference on Computer Vision (ECCV), 2008. 48, 81, 82, 84, 88, 89, 117, 119, 120

[Tsai & Fu 1979] Wen-Hsiang Tsai and King-Sun Fu. *Error-correcting isomorphisms of attributed relational graphs for pattern analysis*. IEEE Transactions on Systems, Man and Cybernetics (TSMC), vol. 9, no. 12, pp. 757 – 768, 1979. 82

[Tupin *et al.* 1998] Florence Tupin, Henri Maitre, Jean-Francois Mangin, Jean-Marie Nicolas and Eugene Pechersky. *Detection of linear features in SAR images: application to road network extraction*. IEEE Transactions on Geoscience and Remote Sensing, vol. 36, no. 2, pp. 434–453, March 1998. 35

[Twining & Taylor 2001] Carole J. Twining and Chris J. Taylor. *Kernel principal component analysis and the construction of non-linear active shape models*. In British Machine Vision Conference (BMVC), 2001. 102

[Umeyama 1988] Shinji Umeyama. *An eigendecomposition approach to weighted graph matching problems*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 10, no. 5, pp. 695–703, 1988. 82

[van Kaick *et al.* 2010] Oliver van Kaick, Hao Zhang, Ghassan Hamarneh and Danial Cohen-Or. *A survey on shape correspondence*. In Proc. of Eurographics State-of-the-art Report, pp. 61–82, 2010. 77, 78, 82

[Vazirani 2001] Vijay V. Vazirani. Approximation Algorithms. Springer, 2001. 37

[Veksler *et al.* 2010] Olga Veksler, Yuri Boykov and Paria Mehrani. *Superpixels and supervoxels in an energy optimization framework*. In European Conference on Computer Vision (ECCV), 2010. 28

[Vicente *et al.* 2008] Sara Vicente, Vladimir Kolmogorov and Carsten Rother. *Graph cut based image segmentation with connectivity priors*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 32

[Vicente *et al.* 2009] Sara Vicente, Vladimir Kolmogorov and Carsten Rother. *Joint optimization of segmentation and appearance models*. In IEEE International Conference on Computer Vision (ICCV), 2009. 49, 119

[Vogiatzis *et al.* 2007] George Vogiatzis, Carlos Hernández Esteban, Philip H. S. Torr and Roberto Cipolla. *Multiview stereo via volumetric Graph-Cuts and occlusion robust photo-consistency*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 29, no. 12, pp. 2241–2246, December 2007. 28

[Vukadinovic & Pantic 2005] Danijela Vukadinovic and Maja Pantic. *Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers*. In IEEE International Conference on Systems, Man and Cybernetics (SMC), 2005. 124, 127

[Wainwright & Jordan 2007] Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends in Machine Learning, vol. 1, no. 1-2, pp. 1–305, 2007. 45

[Wainwright *et al.* 2004] Martin J. Wainwright, Tommi Jaakkola and Alan Willsky. *Tree consistency and bounds on the performance of the max-product algorithm and its generalizations*. Statistics and Computing, vol. 14, no. 2, pp. 143–166, April 2004. 40

[Wainwright *et al.* 2005] Martin J. Wainwright, Tommi S. Jaakkola and Alan S. Willsky. *MAP estimation via agreement on trees: Message-passing and linear programming*. IEEE Transactions on Information Theory, vol. 51, no. 11, pp. 3697–3717, November 2005. 22, 27, 28, 35, 36, 41, 46, 47, 120

[Walker & Herman 1988] Ellen Lowenfeld Walker and Martin Herman. *Geometric reasoning for constructing 3D scene descriptions from images*. Artificial Intelligence, vol. 37, no. 1-3, pp. 275–290, 1988. 105

[Wang & Adelson 1994] John Y. A. Wang and Edward H. Adelson. *Representing moving images with layers*. IEEE Transactions on Image Processing (TIP), vol. 3, no. 5, pp. 625–638, 1994. 56

[Wang *et al.* 2005] Yang Wang, Mohit Gupta, Song Zhang, Sen Wang, Xianfeng Gu, Dimitris Samaras and Peisen Huang. *High Resolution Tracking of Non-Rigid 3D Motion of Densely Sampled Data Using Harmonic Maps*. In IEEE International Conference on Computer Vision (ICCV), 2005. 77, 94

[Wang *et al.* 2007] Sen Wang, Yang Wang, Miao Jin, Xianfeng David Gu and Dimitris Samaras. *Conformal Geometry and Its Applications on 3D Shape Matching, Recognition, and Stitching*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 29, no. 7, pp. 1209–1220, 2007. 79, 80, 95, 96

[Wang *et al.* 2009] Chaohui Wang, Martin de La Gorce and Nikos Paragios. *Segmentation, Ordering and Multi-object Tracking Using Graphical Models*. In IEEE International Conference on Computer Vision (ICCV), 2009.

[Wang *et al.* 2010] Chaohui Wang, Olivier Teboul, Fabrice Michel, Salma Essafi and Nikos Paragios. *3D Knowledge-Based Segmentation Using Pose-Invariant Higher-Order Graphs*. In International Conference, Medical Image Computing and Computer Assisted Intervention (MICCAI), 2010.

[Wang *et al.* 2011a] Chaohui Wang, Haithem Boussaid, Loic Simon, Jean-Yves Lazennec and Nikos Paragios. *Pose-invariant 3D Proximal Femur Estimation through Bi-Planar Image Segmentation with Hierarchical Higher-Order Graph-based Priors*. In International Conference, Medical Image Computing and Computer Assisted Intervention (MICCAI), 2011. 126

[Wang *et al.* 2011b] Chaohui Wang, Michael M. Bronstein, Alexander M. Bronstein and Nikos Paragios. *Discrete Minimum Distortion Correspondence Problems for Non-rigid Shape Matching*. In International Conference on Scale Space and Variational Methods in Computer Vision (SSVM), 2011.

[Wang *et al.* 2011c] Chaohui Wang, Yun Zeng, Loic Simon, Ioannis Kakadiaris, Dimitris Samaras and Nikos Paragios. *Viewpoint Invariant 3D Landmark Model Inference from Monocular 2D Images Using Higher-Order Priors*. In IEEE International Conference on Computer Vision (ICCV), 2011.

[Weiss & Freeman 2001] Yair Weiss and William T. Freeman. *On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs*. IEEE Transactions on Information Theory, vol. 47, no. 2, pp. 736–744, 2001. 35, 40

[Werner 2007] Tomás Werner. *A linear programming approach to max-sum problem: a review*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 29, no. 7, pp. 1165–1179, July 2007. 35, 46

[Werner 2008] Tomás Werner. *High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (MAP-MRF)*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 51

[Werner 2010] Tomás Werner. *Revisiting the linear programming relaxation approach to Gibbs energy minimization and weighted constraint satisfaction*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 32, no. 8, pp. 1474–1488, August 2010. 46, 51

[Winn & Blake 2004] John Winn and Andrew Blake. *Generative Affine Localisation and Tracking*. In Advances in Neural Information Processing Systems (NIPS), 2004. 56

[Winn & Shotton 2006] John Winn and Jamie Shotton. *The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. 109

[Woodford *et al.* 2009] Oliver J. Woodford, Philip H. S. Torr, Ian D. Reid and Andrew W. Fitzgibbon. *Global Stereo Reconstruction under Second-Order Smoothness Priors*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 31, no. 12, pp. 2115–2128, 2009. 31

[Wu *et al.* 2002] Wei Wu, Michael J. Black, Yun Gao, Elie Bienenstock, M. Serruya, A. Shaikhouni and John P. Donoghue. *Neural Decoding of Cursor Motion using a Kalman Filter*. In Advances in Neural Information Processing Systems (NIPS), 2002. 24

[Xiang *et al.* 2011] Bo Xiang, Chaohui Wang, Jean-Francois Deux, Alain Rahmouni and Nikos Paragios. *Tagged Cardiac MR Image Segmentation Using Boundary & Regional-Support and Graph-based Deformable Priors*. In IEEE International Symposium on Biomedical Imaging (ISBI), 2011. 126

[Yang *et al.* 2005] T. Yang, S. Z. Li, Q. Pan and J. Li. *Real-Time Multiple Objects Tracking with Occlusion Handling in Dynamic Scenes*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005. 55, 56

[Yanover *et al.* 2006] Chen Yanover, Talya Meltzer and Yair Weiss. *Linear Programming Relaxations and Belief Propagation-An Empirical Study*. The Journal of Machine Learning Research, vol. 7, pp. 1887–1907, 2006. 46

[Yedidia *et al.* 2003] Jonathan S. Yedidia, William T. Freeman and Yair Weiss. *Understanding Belief Propagation and its Generalizations*. In Exploring artificial intelligence in the new millennium, pp. 239–269. Morgan Kaufmann, 2003. 30, 34, 40

[Yin *et al.* 2006] L. Yin, X. Wei, Y. Sun, J. Wang and M.J. Rosato. *A 3D facial expression database for facial behavior research*. In IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2006. 122

[Yin *et al.* 2008] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm and Michael Reale. *A High-Resolution 3D Dynamic Facial Expression Database*. In IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2008. 122

[Zaharescu *et al.* 2009] A. Zaharescu, E. Boyer, K. Varanasi and R. Horaud. *Surface Feature Detection and Description with Applications to Mesh Matching*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 86

[Zass & Shashua 2008] Ron Zass and Amnon Shashua. *Probabilistic graph and hypergraph matching*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 82

[Zeng *et al.* 2008] Wei Zeng, Yun Zeng, Yang Wang, Xiaotian Yin, Xianfeng Gu and Dimitris Samaras. *3D Non-rigid Surface Matching and Registration Based on Holomorphic Differentials*. In European Conference on Computer Vision (ECCV), 2008. 79, 80

[Zeng *et al.* 2010] Yun Zeng, Chaohui Wang, Yang Wang, Xianfeng Gu, Dimitris Samaras and Nikos Paragios. *Dense non-rigid surface registration using high-order graph matching*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[Zeng *et al.* 2011] Yun Zeng, Chaohui Wang, Yang Wang, Xianfeng Gu, Dimitris Samaras and Nikos Paragios. *Intrinsic Dense 3D Surface Tracking*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[Zhang & Hebert 1999] Dongmei Zhang and Martial Hebert. *Harmonic Maps and Their Applications in Surface Matching*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1999. 79, 80

[Zhang & Ji 2005] Yongmian Zhang and Qiang Ji. *Active and dynamic information fusion for facial expression understanding from image sequences.* IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 27, no. 5, pp. 699–714, May 2005. 24

[Zhang *et al.* 2004] Li Zhang, Noah Snavely, Brian Curless and Steven M. Seitz. *Spacetime faces: high resolution capture for modeling and animation*. ACM Transactions on Graphics (TOG), vol. 23, no. 3, pp. 548–558, 2004. 77

[Zhu & Yuille 1996] Song Chun Zhu and Alan Yuille. *Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multiband Image Segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 18, no. 9, pp. 884–900, 1996. 32