Contribution à l'étude et à la mise en œuvre d'un système de perception bio-inspiré basé sur l'attention visuelle et auditive

Jingyu Wang

UNIVERSITÉ PARIS-EST

西北工業大學
NORTHWESTERN POLYTECHNICAL UNIVERSITY

Thesis

Presented to obtain the title of
DOCTEUR DE L'UNIVERSITÉ PARIS-EST
DOCTOR OF ENGINEERING

Specialization: Signal and Image Processing

By: **Jingyu WANG**

Contribution to study and implementation of a bio-inspired perception system based on visual and auditory attention

Defended on 09 January 2015 in presence of commission composed by

Dr. HDR	Lucile	ROSSI	Reviewer / UMR SPE CNRS - University of Corsica
Prof.	Xinbo	GAO	Reviewer / Xidian University
Prof.	Xuelong	LI	Examiner / Chinese Academy of Science / Northwestern Polytechnical University
Dr.	Christophe	SABOURIN	Examiner / LISSI - University PARIS-EST Créteil
Prof.	Ke	ZHANG	Co-supervisor / Northwestern Polytechnical University
Prof.	Kurosh	MADANI	Co-supervisor / LISSI - University PARIS-EST Créteil

UNIVERSITÉ PARIS-EST

西北工業大學
NORTHWESTERN POLYTECHNICAL UNIVERSITY

Thèse

Présentée pour l'obtention des titres de
DOCTEUR DE L'UNIVERSITÉ PARIS-EST
DOCTOR OF ENGINEERING

Spécialité: Traitement du Signal et des Images

Par : **Jingyu WANG**

Contribution à l'étude et à la mise en œuvre d'un système de perception bio-inspiré basé sur l'attention visuelle et auditive

Soutenue publiquement le 09 janvier 2015 devant la commission d'examen composée de

Dr. HDR	Lucile	ROSSI	Rapporteur / UMR SPE CNRS - Université de Corse
Prof.	Xinbo	GAO	Rapporteur / Xidian University
Prof.	Xuelong	LI	Examinateur / Chinese Academy of Science / Northwestern Polytechnical University
Dr.	Christophe	SABOURIN	Examinateur / LISSI - Université PARIS-EST Créteil
Prof.	Ke	ZHANG	Codirecteur de thèse / Northwestern Polytechnical University
Prof.	Kurosh	MADANI	Codirecteur de thèse / LISSI - Université PARIS-EST Créteil

Acknowledgement

First and foremost I would like to express my deepest gratitude to my advisor from Université Paris-Est Créteil, Professor Kurosh Madani, for his professional instruction and invaluable support to my doctoral research. I would like to thank him for encouraging my research with his unsurpassed knowledge as well as the visionary thoughts. It has been a great honor for me to conduct my work under his supervision and to be one of his PhD students.

I would like to give my sincerest appreciation to my advisor from Northwestern Polytechnical University, Professor Ke Zhang, for the continuous support and help of my PhD research. I am grateful to him wholeheartedly, not only for his tremendous academic advice, but also for providing me with so many precious opportunities and unconditional trust.

I would like to extend my special thanks to my tutor from Université Paris-Est Créteil, Dr. Christophe Sabourin, for his generous support and great patience. I am very grateful to him for his scientific experience and especially his guidance which helped me in all the time of research and writing of this thesis.

I would like to thank my committee members from French and Chinese sides, Prof. Xuelong Li from Northwestern Polytecnical University, Prof. Xinbo Gao from Xidian University, Dr. HDR Lucile Rossi from Université de Corse, for serving as my committee members. I want to thank you for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, many thanks to you.

My sincere thanks also should given to Prof. Yanjun Li, Prof. Jianguo Huang, Prof. Jingdong Chen, Prof. Yanning Zhang, Prof. Yixin Yang, Prof. Xiansheng Qin, Prof. Geng Liu, Prof. Licheng Jiao, Prof. Shuyuan Yang, Prof. Maoguo Gong, for their scientific suggestions, professional advices, helps and encouragements on my PhD research work.

I am also very grateful to the members and ex-members of LISSI and my Chinese laboratory including: Dr. Weiwei Yu, Dr. Ting Wang, Dr. Dominik Maximilián Ramík, Ouerdia Megherbi, Assia Aziez, Dr. Meibo Lv, Dr. Pei Wang, Dr. Minghuan Zhang, Dr. Yang Chong, Shun Xue, Jianfeng Cui, Cong Nie, Wenjun Yang, Xue Fang, Hao Gao, Zhiguo Han, Haixu Jiang, Minghu Tan, Di Jiang, Jing Zhang, Ye Zhang, Ruige Zhang, Yali Qin, Yu Su, Haoyu Li, Zhichen Tan, for their generous supports and kind helps to my research work and doctoral life. I am truly grateful for their friendship.

I want to thank Dr. Guokang Zhu, Dr. Huashan Feng, Dr. Chaoya Liu, Dr. Jun Li, Dr. Yanfeng Wang, Dr. Yaozhen Wu, Dr. Mingxing Xu, Dr. Qiong Nie, Zhennan Guo, for their assistances to my doctoral study in France and China.

At last, but by no means least, I would like to especially express my most sincere thanks and heartfelt gratitude to my dear parents: my father Runxiao Wang and my mother Qunxiu Yao, for giving birth to me and sacrificing their lives to raise me with generous love and care. Words can not express how grateful I am to them. I thank all the members of my entire family and their love is what sustained me thus far.

I dedicate this thesis
to my most beloved parents
and to all my family members
for their love

Table of Contents

List of Figures

List of Tables

List of Symbols

s	Sound signal
s_j	The j^{th} frame of sound signal
$E_j(s_j)$	Short-term Shannon entropy of j^{th} frame
$E_f(s)$	Short-term Shannon entropy of sound signal s
$\mathrm{G}(s_t)$	Global saliency estimation of sound
σ_s	Threshold which determines the global saliency
$\omega(t)$	Window function for smoothing the discrete value
BGN_s	Background noise estimation
$0_nG(s_t)$	Number of zero in $G(s_t)$
θ_s	Threshold of global saliency indicated by the SSE
$mel(f)$	Mel-scale frequency
t_{IOR}	Minimum time interval of inhibitory effect in temporal saliency
M_p	MFCC value of p point
M_{p+1}	MFCC value of $p+1$ point
t	Time length between p and $p+1$ salient point
α_s, β_s	Adjustment coefficients
$M_{IOR+,p+1}$	Upper limit of the expectation value for $p+1$ point
$M_{IOR-,p}$	Lower limit of the expectation value for p point
$M_{IOR-,p+1}$	Lower limit of the expectation value for $p+1$ point
f_q	Location of the q salient point in frequency axis
f_{Pmean}	Location of frequency's mean value point
$S_{PSD,q}$	True saliency property of q point
P_q	PSD value of the q salient point
P_{mean}	The mean value of PSD

$P_{local\min,j}$	Local minimum value in the j^{th} frequency band
$P_{local\min,j+1}$	Local minimum value in the $j+1^{th}$ frequency band
ζ	Saliency distance of PSD
$\forall$	Arbitrary sign
I_s	Spectrogram of sound
I_g	Log scale spectrogram
$I_{r\text{-}g}$	Saliency from red-green channel
$I_{g(red)}$	Red channel information
$I_{g(green)}$	Green channel information
$I_{red\text{-}green}$	Image saliency of red-green channel
RGB_r, RGB_g	Red and green channel of RGB color space
$\{S_{SSE}\}$	A set of salient segments indicated by $G(s_t)\neq 0$
$\{S_{Mp}\}$	A set contains temporal locations of salient signals
$\{S_{P>Pmean}\}$	A set contains salient frequency components which PSD values are greater than P_{mean}
$\{S_{PSD,q}\}$	A set contains salient components in frequency domain
$\cap$	Intersection operation
$\cup$	Union operation
$\varnothing$	Null set
$\oplus$	Logic plus operation
$\|\|$	Logic and operation
S_{image}	Image saliency feature
$S_{spectral\text{-}temporal}$	Spectral and temporal saliency feature
S_{af}	Final auditory saliency feature
A	A fuzzy set
X_f	Collection of all the objects
x_f	Generally element of a fuzzy set
$\mu_A(x_f)$	Membership function of x in a fuzzy set A
$x_{f1}, \ldots x_{fn}$	Infinite elements of X_f

$\mu_A(x_{fi})$, $\mu_A(x_{fi})/x_{fi}$	Membership grade of x_{fi}
P_s	Principal spectral component of PSD
λ_p	A adjustment factor that determines the energy ratio of principal spectral component
Δ’s	First derivatives
$\Delta\Delta$’s	Second derivatives
d_i	Delta coefficient of frame i
c_{n+i}, c_{n-i}	Basic coefficients of MFCC
$g_c(\cdot)$, $g_s(\cdot)$	Cosine and sine function with a Gaussian window
ω_0	The frequency at which Gabor filter has maximum response
σ_g	Spread of Gaussian window
$g_c(\cdot)$, $g_s(\cdot)$	Gabor filters of even (e.g. cosine) and odd (e.g. sine) waves
$\left(\omega_{x_0}, \omega_{y_0}\right)$	Centre frequency of x and y axis
(σ_g, σ_g)	The lengths of Gaussian window in x and y axis
$g_{complex}$	Complex version of Gabor filter
g_{real}	Real component of Gabor filter
$g_{imaginary}$	Imaginary component of Gabor filter
λ	The wavelength of the sinusoidal wave
θ_g	Orientation of the parallel stripe of each Gabor function
ψ	Phase offset
σ_{sg}	Standard deviation of Gabor function
γ	Spatial aspect ratio
$\psi_{u,v}$	Kernel of Gabor function
$k_{u,v}$	Wave vector
u, v	Orientation and scale of the Gabor kernel
x_l, y_l	Coordinates of pixel location
$\|\cdot\|$	Norm operator

σ_{gr}	Ratio of the Gaussian window width to wavelength
k_{max}	Maximum frequency
f_{sf}	Spacing factor between kernels in frequency domain
$I(z)$	Image
$G_{u,v}(z)$	Gabor feature of image in orientation u and scale v
$*$	Convolution operator
F	Fourier transform
F^{-1}	Inverse Fourier transform
$M_{u,v}(z)$	Magnitude of Gabor features
$\theta_{u,v}(z)$	Phase of Gabor features
U_i	Element uniqueness
c_i, c_j	CIELab colors of segment i and j
p_i, p_j	Positions of segment i and j
$\omega(p_i, p_j)$, $\omega_{ij}^{(p)}$	Weight relates to the distance between segments
σ_p	Operation range
Z_i	Normalization factor
D_i	Color uniqueness
$\omega(c_i, c_j)$, $\omega_{ij}^{(c)}$	Color similarity of segments i and j
μ_i	Weighted mean position of color c_i
Sv_i	Saliency value for each element
k_s	Scaling factor
$\hat{S}v_i$	Final saliency of a pixel in an image
α_v, β_v	Parameters controlling the sensitivity to color and position
x	Column signal
$\Re^n$, $\Re^m$, $\Re^{n\times K}$	Euclidean space of n, m, $n\times K$ dimension
K	Number of atoms

D	Overcomplete dictionary
$\|\cdot\|_0$	The l^0-norm operation
ε	Error tolerance
$\|\cdot\|_1$	The l^1-norm operation
α	Sparse coefficient
γ_c	Constant
α_i	The i^{th} sparse coefficient
$D_{i,j}$	Element of i^{th} row and j^{th} column in dictionary D
L	Lagrangian dual
l_j	Dual variable
$(\cdot)^{\mathrm{T}}$	Transposition operation
Λ	The diagonal element
e_i	The i^{th} unit vector
∂	First-order derivative
∂^2	Second-order derivative
x_i	Feature vector of the i^{th} image patch
$\hat{x}$	Reconstructed feature vector
$\|\cdot\|_2^2$, $\|\cdot\|_E$	Euclidean distance
ε_{fo}	A set consists of reconstruction errors
$\varepsilon_i^{s_j}$	Reconstruction error of i^{th} local image patch in s_j scale
ρ_{s_j}	Erro threshold
C_{object}	Classification results of foreground objects
$O_{classify}\{\cdot\}$	Object classification operator
f_{vs}	Visual saliency feature
$f_{objectness}$	Objectness feature
x_v, x_v'	Different input image feature vector
$M(\cdot)$	Mapping operation

C	A set of codebook
C_i	Codeword of a codebook
$d(\cdot)$	Calculation operator of distortion
Q_i	Unique Voronoi cell
□	Mapping function of SVM
C_s	Positive regularization parameter
p	Multinomial distribution
T	Number of dimension for Dirichlet distribution
α_j	A set of hyperparameter
$V_{i,j}$	Visual information of i^{th} object in j^{th} environment
$V_{i,j}^{P}$	Probability of $V_{i,j}$
$S_{i,j}$	Acoustic information of i^{th} object in j^{th} environment
$S_{i,j}^{P}$	Probability of $S_{i,j}$
V_j, S_j	Sets of appearance probability of j^{th} visual and acoustic information
$P_{V_j}^{i}, P_{S_j}^{i}$	Probability of V_j, S_j *in* i^{th} *IPM*
Obj_k	The k^{th} object that could appeared in the environment
Obj_l	The *IPM* of l^{th} structural salient object
$P_{nobj}^{i,l}$	Probability of Obj_l in i^{th} environment scene
$\cup$	Assemble operation
N_v	Number of visual objects
N_a	Number of acoustic objects
τ_p	Probability threshold

Chapter 1. General Introduction

1.1. Foreword

Since the birth of mankind, the exploration of surrounding environment has not been stopped, even for one second. So far, the living space of human has been tremendously enlarged and so as the information of where we lived. During the process of getting knowledge of the environment, perception is a fundamental ability of human beings which describe the process of knowing the surrounding information.

The perception ability of human beings could be seen as a complex recognition system by using the information of surrounding environment from both human vision system (HVS) and human hearing system (HHS). These two systems represent two major perception channels of our body. In general, the human perception system is very intelligent, fast and could be able to perceive environment along with the existed objects which endowed with heterogeneous information at the same time. However, modern artificial machines that frequently been used nowadays still can not combine sound and image data to merge into a unified paradigm of information like human beings, because the fusion of sound and image information is such a great challenge to artificial intelligence due to the difficulties of processing and fusing heterogeneous information in a bio-inspired way. Therefore, novel processing and fusion techniques regarding sound and image information should be researched in order to form higher level information for intelligent perception.

Researches from neuroscience have shown that human brain is a complex and sophisticated network which comprised by tremendous amount of neurons. All the information achieved by our eyes and ears will be processed by this neural network at the same time no matter the information is either homogeneous or heterogeneous. Meanwhile, there existed a selected attention perception and learning principle called awareness when we come cross some salient objects in either sound or image domain. Though quantity of researches have been conducted and proved to be effective in obtaining salient information from sound and image signals, their applications in environment information acquisition still faces limitations and shortcomings because of the complexity of real environment. Consequently, development should be made to provide better awareness ability for artificial machines. It should be emphasize that, saliency principle is considered as a key component of awareness ability as it allows the perception in both visual and auditory channels to be carried out in an efficient and effective way. Thus, it is used not only as a universal rule of perception but also the fundamental basis of information processing through all the chapters that related to the information acquisition in this thesis, based on which novel processing methods are researched and combined to yield a better performance in information acquisition.

Regarding another critical research issue of information fusion in this thesis, the fusion work could be done in data level, feature level and decision level, theoretically. In most researches, as the fusion of different kinds of information is considered and conducted between homogeneous information, such as visible image and infrared image, ultra sonic data and radar data, etc., the fusion processing could be done in all the three levels. However, data level fusion can hardly be possible to be carried out for sound and image information, because they are heterogeneous information and have no correlation in data level. That is to say, the information in each pixel of image domain has no corresponding relationship to the sound data that referred to the same object at any time, which means that the value of image pixel has no perceptual connection with the value from acoustic domain. Therefore, most information fusion methods, probably all of them could not be used in data level fusion due to these natural difficulties generated by the mechanism of how humans collect information,. Surprisingly enough, human beings can process heterogeneous information of image and sound synchronously in the exploration and perception of environment in our everyday life. This complex perception procedure is done spontaneously because human brain has been naturally trained for years to accomplish this sophisticated work based on heterogeneous information. Thus, this human-like perception process is the research goal and the main technique for realizing artificial awareness.

1.2. Biological Background

To help understanding the biological inspirations of my research work, some relevant background will be given from the perspectives of human perception ability and characteristics. Since the artificial awareness for intelligent machine can be seen as the simulation of human perception ability in which saliency principle is a baseline technique for information acquisition, biological inspirations obtained from human perception system and selective attention mechanism are illustrated in order to give an overview of how I process the heterogeneous information acquisition of environment in a human-like way and why I use saliency as the general processing rule to realize the artificial awareness.

1.2.1. Human Perception System

From a semantics point of view, the word "perception" is defined by dictionary[1] as "the ability to see, hear, or become aware of something through the senses", while in the theory of psychology, the perception is the organization, identification, and interpretation of sensory information in order to represent and understand the environment. To explain the fundamental truth of perception, psychologists have done a lot of researches to reveal the inner activity of human brain when perception process is carried out. Researches have shown that the perception process is actually a series of stimuli, transduction and re-creation activities inside the nervous system of human

[1] Oxford Dictionary of English, (Stevenson, 2010)

beings which involved sensory organs, neurons and brain, respectively. In biology research, the sensory organs are also called the receptors, the responsibility of which is to receive the stimulus energy from either surrounding environment or the self-body of human themselves. The receptors that human beings mostly used are eyes, ears, skin, muscles, nose and tongue, all of which can be divided into two major categories: environment receptors and self-body receptors, thus, the human perception system can also be divided into two subsystems: the environment perception system and the self perception system.

In the environment perception system of human body, eyes and ears are the two main sensory organs which will convert the input energy of the environmental stimuli (e.g., the light energy that strikes the retina of our eyes and the pressure wave that generate fluid-borne vibrations in the organ of Corti of the ear) into signals of nerve impulses which generated by different kinds of nerve cells, and the visual information and auditory information will be gained after the bio-electrical signals are transmitted and processed by our brain, or rather, the certain parts of cerebral cortex which are visual cortex and temporal lobe, respectively.

It should be noticed that, these two receptors are receiving environmental stimuli at all time in most cases, even when we are fall asleep, this means that the brain of human beings must process the heterogeneous information from both vision system and auditory system at the same time. Researches in cognitive neuroscience and cognitive psychology have shown that, the perception process of human involves two different patterns: top-down and bottom-up (Frintrop et al., 2010). These two patterns described two different kinds of processing sequences, or flow (Palmer et al., 1981), of information in human brain, where “down” means sensory input (i.e. data level information) and “up” refer to the higher cognitive process (i.e. knowledge level information). The top-down processing pattern is known as executive attention or the conceptually-driven processing, this pattern is driven by priori knowledge and internal hypotheses to form a certain expectation to direct the perception process. In top-down pattern, the expectation of perception will make the most related low level information a higher priority to be processed, while the most unrelated information is ignored so that the perception resources (e.g., memory, compute) could be used more efficiently (Kim et al., 2013). On the contrary, the bottom-up processing pattern is known as the data-driven processing or stimulus-driven attention. Generally, it is the perception processing that use low-level information to acquire higher-level information, such as knowledge and explanation. In bottom-up pattern, the perception processing is driven by salient stimuli which generated by the salient features or information of the objects to form low-level information from environment, then higher-level information such as decision knowledge could be obtained after processed by brain (Liu et al., 2009). Therefore, it is believed that selective attention mechanism directs the bottom-up perception processing, and this perception pattern is very fast and pre-conscious because it is a non-volitional way of perception outside the conscious awareness (Goldstein, 2013).

1.2.2. Selective Attention Mechanism

Selective attention refers to the processes that only particular stimulus will be selected and reacted for higher-level information processing while the irrelevant or unexpected information be suppressed simultaneously (Steven et al., 2012). Supported by the researches of cognitive neuroscience, selective attention mechanism is proved to be existed in top-down pattern where salient information or stimulus guiding the attention allocation of human perception system. Meanwhile, an unconscious but similar one exists in bottom-up patterns as well, where the goals or expectations of an individual will lead to attention (Yantis and Serences, 2003; Buschman and Miller, 2007). Cognitively speaking, as an evolution result of millions of years, selective attention mechanism could help mankind to perceive the surrounding environment in a more efficient way as it could direct the brain to selective concentrate on one aspect of the environment or objects while ignore the others in order to selectively allocate the perception resources and make the perception processing more quickly (Anderson, 2005).

Early researches from cognitive psychology have developed several models, such as filter model (Cherry, 1953; Broadbent, 1958), attenuation model (Treisman, 1960) and response selection model (Deutsch and Deutsch, 1963), to explain how human attention could be selective in either visual or auditory perception systems. The main precondition of these models is the limitation of information processing capability in human brain, which causes the existence of such kind of models to help human beings to select or filter only interested or salient low-level information to be processed in higher-level processing so as to avoid the "bottleneck" of cognitive processing capability in cognitive psychology research. As the filter model and attenuation model share the same motivation, location of filter and property in cognitive theory, they are also called the perception selection model of attention. Supported by experimental results, perception selection model and response selection model could partly explain the selective mechanism of attention.

However, reliable evidences still could not be provided by researches or experimental results to fully support above models because the actual existence of such filter or bottleneck in human brain could not be specified by neuroscience research. Thus, other researchers have proposed an energy distribution based theory called central capacity theory which described selective attention as the representation of energy distribution system in human brain (Kahneman, 1973). The central capacity theory state that the ability for human beings to response to a stimulus from external is based on the amount of cognitive processing resources that each person has, where the allocation of resources decide what stimulus should be responded by perception system and restriction will be executed to emphasize and enlarge the most salient stimuli while ignore the less interested ones. According to this theory, spotlight model (Posner et al., 1980) and zoom-lens model have been proposed to indicate how visual attention mechanism operates.

Generally speaking, despite the differentials among all these theoretic models in cognitive psychology research, many researchers from computer science and artificial intelligence have gradually accepted that selective attention mechanism enable human beings to focus on and response to the most salient stimulus occurred in surrounding environment in an unconscious but fast way, and it could also be the most efficient approach in searching the expectation objects when we explorer the natural world since the more concentrative the perception system is, the more detailed information we will have. In the biology research work of (Michael and Gálvez-García, 2011) a modified visual search paradigm which slightly different in size is used to conduct a series of behavioral experiments and the results show that visual attention is captured by the most salient as well as the least salient item and progress in a salience-based fashion until a target is found. (Kayser et al., 2005) presented an auditory saliency map based auditory attention model for auditory scene perception, in which the sound wave is converted to a visualized time-frequency representation. The auditory features from different scales are extracted in parallel by using saliency as a common principle and multiple saliency maps for individual features are finally added to yield the saliency map for sound signal. The structure of this model is identical to that of successfully used visual saliency maps, and the experimental results demonstrate that the mechanisms extracting conspicuous events from a sensory representation are similar in auditory and visual pathways. Moreover, (Fritz et al., 2007) described the inter-modal and cross-modal interactions between auditory and visual attention from a neurobiology point of view, suggested that auditory and visual attention modalities share similarities in both bottom-up and top-down selective attention frameworks where image-based saliency cues is thought to be operated as the main mechanism.

Hence, the saliency principle based technology or algorithm could be a possible way to realize the selective attention mechanism in computational term. Meanwhile, in both “top-down” and “bottom-up” attention patterns, selective attention mechanism is also the necessity and gateway to the awareness ability exist in either visual or auditory perception system which supported by the results of quantity of experiments which conducted by neuropsychologists in (Dehaene and Naccache, 2001; Dehaene et al., 2006; Eriksson et al., 2008; Mesgarani and Chang, 2012; Caporello Bluvas and

Gentner, 2013). Therefore, saliency principle based information acquisition technique is considered to be a practical way of processing heterogeneous information similar to human perception system.

1.3. Motivation and Objectives

The motivation of my thesis comes from the question that can machine be more intelligent, or even be conscious. Since the day computer was invented, it has been the ultimate goal of artificial intelligence to develop a machine that has the characteristic of consciousness. In recent years, many research works like (Edelman and Tononi, 2000; Jennings, 2000; Aleksander, 2001; Baars, 2002; Kuipers, 2005; Manzotti, 2006; Adami, 2006; Chella and Manzotti, 2007a; Edelman and Tonomi, 2013) has been put out to explore the hypothesis of designing and implementing models for artificial consciousness (Buttazzo and Manzotti, 2008). On one hand there is hope of being able to design a model for consciousness (McDermott, 2007), on the other hand the actual implementations of such models could be helpful for understanding consciousness (Edelman and Tononi, 2000). According to (Holland, 2003) and (Seth, 2009), the definition of artificial consciousness could be divided into two major aspects, which are:

1) *Weak Artificial Consciousness: design and construction of machine that simulates consciousness or cognitive processes usually correlated with consciousness.*

2) *Strong Artificial Consciousness: design and construction of conscious machines.*

It would be easier for the majority of artificial intelligence researchers to accept the first definition because an artificial conscious being would be a being which appears to be conscious when acts and behaves as a conscious human being (Manzotti and Tagliasco, 2008). Many researchers have developed several computational models to realize the goal of artificial consciousness in order to make the machine, such as computer, act more like human beings and therefore, be conscious. However, most of them are more likely to be a metaphorical description than an implementable model in computational terms (Arrabales et al., 2009). To be more specific, it is extremely difficult for engineers and scientists to design an intelligent artificial artifact that could be able to make experience, because the 'experience' derives from examining the behavior of an entire human being with all his history and his capacity of communicating, thus there is no practical way for robot to gain such 'experience' by itself autonomously only if human beings teach it or make the robot be aware of such kind of experience. Therefore, indicated by the work of (Chella and Manzotti, 2007b), although a lot more advanced than in the past, artificial intelligent systems such as robots, still fall short of human agents, and the characteristic of consciousness could be the missing step in the ladder from current artificial robots to human-like robots.

Considering the relationship between attention and awareness, it could be either close functional (Kanwisher, 2001), or complex (Koivisto et al., 2009). As the representation of human consciousness, the awareness ability of human is defined as the state or ability to perceive, to feel, or to be conscious of events, objects, or sensory patterns. Since current researches still could not develop computational models to mimic the characteristic of artificial consciousness which is the updated version of artificial intelligence, the ability of artificial awareness could be the first step for researchers to improve the performance of autonomous characteristic in perception process of robot, which is the symbol of artificial intelligence. In (Fuertes and Russ, 2002) a perceptive awareness model including three specific functions which defined to be performed in order to reach perceptive awareness for automatic systems is presented, in which the functions refer to the perception and recognition processes of the current situation on one hand, and the selection and evaluation of proper response on the other. Meanwhile, perception data from both microphone and cameras are prior considered in this model and data combination function is also used to result in a better and more accurate perception. The demonstration of this model indicates that environmental perception is the essential aspect for automatic system, or rather, the machine to obtain the perceptive awareness ability. Therefore, to achieve the goal of intelligent machines, adding artificial conscious awareness or information processing capabilities associated with the conscious mind, would be an interesting way, even a door to more powerful and general artificial intelligence technology (Reggia, 2013). Consequently, inspired by the salient stimuli based selective attention mechanism, saliency principle based artificial awareness ability in exploring the environment of machines is the priori consideration and also a practical approach to provide machines with intelligent and autonomous way of perceiving the heterogeneous information from the same object or event.

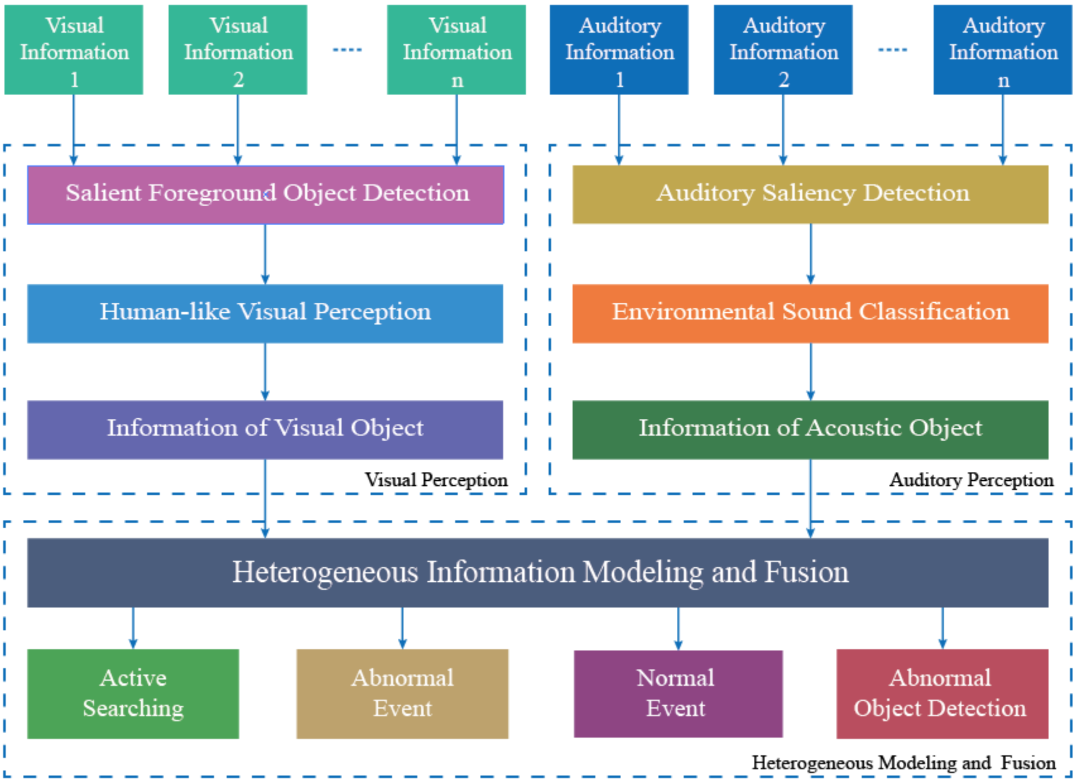
To practically provide machines with artificial awareness ability that can be computational realizable based on heterogeneous information, the salient information acquired from sound and image signals should be successfully obtained and processed to form knowledge individually, which then will be combined together to achieve the higher level information (i.e. knowledge) in a heterogeneous and autonomous way for the final realization of artificial awareness for intelligent perception. Therefore, the research objectives of my work accomplished in this thesis can be summarized in three aspects:

- Investigate the state of the art of auditory saliency detection research and audio classification that relates to the topic of environment sound. Develop novel auditory saliency detection method for salient environment sound detection, explore and design proper features to realize the classification of environment sound for acoustic perception.
- Research and realize the state of the art of visual saliency detection as well as object classification, explore the detection of foreground environmental object. Perform novel salient foreground object detection on both visually salient and semantically salient environmental objects for classification to obtain the knowledge required by visual awareness.
- Establish a novel heterogeneous information fusion framework for artificial perception by regarding heterogeneous information of sound and image as complementary and equally important representation of environment object

or event. Develop an intelligent perception technique that could provide comprehensive understanding of environment object and event for complex environment.

1.4. Contribution

The general overview of proposed approach is illustrated in Figure 1, in which the functions of three major process modules are presented.



Visual Information 1
Visual Information 2
·····
Visual Information n
Auditory Information 1
Auditory Information 2
·····
Auditory Information n
Salient Foreground Object Detection
Human-like Visual Perception
Information of Visual Object
Visual Perception
Auditory Saliency Detection
Environmental Sound Classification
Information of Acoustic Object
Auditory Perception
Heterogeneous Information Modeling and Fusion
Active Searching
Abnormal Event
Normal Event
Abnormal Object Detection
Heterogeneous Modeling and Fusion

Figure 1: The general overview of proposed approach.

It has been demonstrated in Figure 1 that, the visual and auditory information of environment are perceived and processed by visual and auditory perception module, respectively. To be specific, in visual perception module, the salient foreground object will be detected by applying objectness and semantic saliency characteristic of object. Then the candidate of visual representation is processed to obtain the information of visual object based on a human-like visual perception mechanism. While in parallel, the salient environmental sound will be detected by using auditory saliency model and classified to achieve the information of acoustic object in auditory perception module. Thereafter, a heterogeneous information modeling and fusion module is deployed, in which the information probability models of object and scene are initially proposed to implement multiple perception tasks in an intelligent way.

The work accomplished in this thesis has achieved several contributions to the research topic of artificial awareness ability based intelligent perception for machines in complex environment:

- First, the state-of-the-art techniques and related researches focus on salient information acquisition from sound and image channels have been studied, which reveal the complexities and difficulties of problems being dealt with in this thesis. The bibliographical studies have demonstrated the limitations and shortcomings in previous researches, leading to the motivation of using saliency property as the outline of potential solutions in effectively dealing with heterogeneous information. It has also been pointed out by the given discussions that, practical paths to realize artificial awareness for machines rely on the effectively acquisition as well as the comprehensively fusion of salient heterogeneous information.
- The second contribution is the proposition and realization of an auditory saliency detection approach, which uses multiple saliency features obtained from both spectral and temporal domain for auditory saliency detection. It should be emphasized that a bio-inspired computational exhibition of return (IOR) model is initially proposed to accurately extract the salient temporal component, which improves the detection performance in dealing with real environment sound when multiple salient sounds existed. In this approach, saliency features from the spatiotemporal channel and the visual channel are combined together by applying a heterogeneous fusion process to yield the final detection result, which makes this approach adequate for use in the

perception of real environment sound that researched in this thesis.

- The third contribution is the design and application of fuzzy vector based acoustic feature for real environment sound classification. By applying the fuzzy representation characteristic of acoustic awareness in spectral domain, the feature is based on the spectral energy distribution of sound and can be seen as a proper representation of spectral property of sound. Regarding the differentials of energy distribution in various kinds of environment sounds, this fuzzy feature of sound can be suitable for use in real environment sound classification and achieve competitive classification results in the combination with other popular temporal features.
- The fourth contribution is the conception and realization of a novel salient foreground object detection approach from visual channel. Different from the traditional notion of saliency which denotes the purely bottom-up based visual saliency property, the concept of saliency in this thesis has been extended to the semantics level as visually salient object in the environment is not permanently the expected object. By this extension, the foreground environment objects can be seen as salient compared with the background, which reveals the objectness characteristic and describes the uniqueness of potential foreground regions. Motivated by the sparse representation theory, the detection of this semantic saliency property is realized based on the reconstruction error of an input image over an overcomplete background dictionary by using the state-of-the-art sparse representation algorithm. The semantic dissimilarity between foreground object and background provides valuable information to the detection of foreground object, which makes it satisfactory for the realization of awareness of environment objects required in the research of this thesis.
- The fifth contribution is the conception and realization of a human-like salient foreground environment object perception approach. This approach combines visual saliency and semantic saliency together in a fusion way to

discriminately form the artificial awareness ability, the general concept of which is to use visual saliency detection as a prior processing to accelerate the perception process and improve the performance of environment object awareness by using the mentioned semantic saliency detection. It allows a more human-like visual perception procedure similar to human awareness characteristic for realizing artificial awareness, which particularly suitable for the perception requirement of intelligent machines researched in this thesis.

- The fifth contribution is the research of proposition and realization of a novel heterogeneous information fusion model which based on the initially proposed information probability model (IPM) of environment object and scene. Motivated by the classical probability topic model which frequently used in semantic analysis of document, the information probability model is designed to be a more comprehensive representation of environment object by considering the heterogeneous information as probabilistic distributed information of object. Thus, the heterogeneous information of an object can be treated as equally important and processed in a unified model, which provides convenience and coherence in dealing with object's heterogeneous information. Moreover, as environment scene is composed of various kinds of environment objects, the scene information probability model (sIPM) is proposed to comprehensively describe the environment event. By extending the general saliency principle to the concept of structural saliency in higher level, the negative information property of an object as well as a scene is initially proposed to provide more vividly understanding of the complex environment. The realization of mentioned models is performed by using a novel fusion framework which can be applied in multiple perception tasks, such as abnormal object detection, normal and abnormal events detection. It enables artificial awareness ability to be achieved in different environments based on heterogeneous information, which makes the intelligent perception a realizable function that has been accomplished to machines in this thesis.

1.5. Organization of Thesis

The content of this thesis composed of five chapters. From the first chapter to the fifth chapter, the reader will be guided from the state of the art of research that relates to my work conducted in this thesis, through the approaches that I proposed as well as realized for providing artificial awareness ability for machines to the final fusion stage for achieving multiple perception tasks in complex environment. The specific content of each chapter can be organized as follows:

Chapter 1 provides the general overview and structure of my thesis, in which the biological background that inspired my work is presented to explain the importance as well as the necessity of using saliency principle as the major criterion. Meanwhile, the motivation and objectives of my work is discussed, to help reader to better understand why I want to achieve artificial awareness for the realization of intelligent perception of machines and what should be achieved towards the research goals. Thereafter, the contribution and organization of my thesis are provided to highlight my work and the structure of this thesis.

Chapter 2 illustrates the state of the art in different research areas that relate to my work. It comprehensively discusses and reviews previous research works regards the awareness and perception of salient information from both visual and auditory channels. To be specific, as current researches are rarely conducted particularly on the topics of artificial awareness and perception, those approaches of visual and auditory saliency detection are presented as potential realization solutions of computational awareness, while the classification and recognition approaches of environment sounds and visual objects are demonstrated as the ways of acquiring information from the heterogeneous signals. Furthermore, the discussion of state of the art in the research of audio-visual fusion brings out the requirement of novel fusion technique regarding heterogeneous information of sound and image. This chapter should give the general overview of state of the art as well as existing techniques in corresponding research areas, based on the limitations and shortcomings of which I further develop my own approaches and accomplish the research objectives in the rest of my thesis.

Chapter 3 deals with the detection and classification issue of salient environment sound. In the auditory saliency detection part, I propose an acoustic saliency detection approach based on the fusion of multiple saliency features from the visual, spectral and temporal domains. To improve the detection accuracy of proposed approach in dealing with real environment sound signals, I initially propose a short-term Shannon entropy method for the estimation of global saliency characteristic of sound and a computational inhibition of return model for validating the temporal saliency. In the classification part of environment sound, I propose a fuzzy vector based feature to represent the spectral energy distribution of environment sound and achieve the information of sound channel by applying the combination of the proposed feature and temporal feature of MFCC. Experiments of both auditory saliency detection and salient sound classification are performed on the real environment sounds to validate the proposed auditory awareness approach.

Chapter 4 focus on the detection and classification of foreground environment objects. Different from the traditional visual saliency detection works, the foreground object is referred as the semantic salient compared to the background. By considering the objectness characteristic, I propose a foreground environment object detection approach based on the dissimilarity between objects and background. The approach is realized by applying sparse representation, where I propose to use the reconstruction error to represent the semantic saliency. To simulate the human perception process, a foreground object detection approach is presented in this chapter, which based on the fusion of visual saliency and semantic saliency. The obtained detection result is used for further classification to yield the visual information of foreground objects for the next step fusion work described in the next chapter. Experiments are carried out on the real world image to verify the proposed visual awareness method.

If the previous two chapters (e.g. Chapter 3 and Chapter 4) could be seen as the computational realization techniques of artificial awareness, Chapter 5 is dedicated to the realization of heterogeneous fusion approach for intelligent perception in complex environment. Inspired by the processing model proposed in the field of semantic analysis, I initially proposed an information probability model of environment object to comprehensively fuse sound and image information. The proposed model regards heterogeneous information as the probabilistic distribution of object's information, thus the importance of sound and image information could be considered equally. As environment scene will consist of multiple objects and vary in different environments, the scene information probability model is presented in this chapter by considering environment scene as a probabilistic distribution of objects. By extending the notion of saliency to structural saliency, the abnormal or semantic distinct object is described as structural salient in a scene. Thus, the negative property of a scene is proposed to represent the structural salient object in opposite to the representive object of a scene. Thereafter, I propose a unified model to combine the models of environment objects and scene together for the understanding and analysis of complex environment. Based on the aforementioned models, the intelligent perception technique is proposed in the format of multiple fusion steps so that different perception tasks can be achieved. By performing the proposed models and technique on several simulated scenarios for validation and the results show that multiple perception tasks could be accomplished in different environments by using heterogeneous information. Moreover, it allows the interaction with human beings and has the capability of dealing with unknown object or event.

The final chapter of this thesis is the General Conclusion, where the reader will be given a summarized conclusion of my research work presented here. Finally, the perspectives of potential future work regards my research accomplished in this thesis are provided.

Chapter 2. Salient Environmental Information Perception - State of the Art

2.1. Introduction

Information from sound and image is heterogeneous according to the differential in generation mechanism and representation domain which leads to heterogeneous features. Though the techniques used in each processing area are well researched and proved to be highly efficient, machines still could not merge these heterogeneous information into a higher level intelligent knowledge for artificial awareness based environment perception ability. It is because that, current approaches are absolutely different in the term of computational model for sound or image signal processing, which makes the heterogeneous features hardly possible to be fused for further fusion work. Meanwhile, the major difficulty occurred in heterogeneous feature acquisition stage is the selection of appropriate feature which share the same feature space that available for sound and image information fusion work.

Since my work is inspired by the previous researches done in the fields of sound and image processing, machine learning and information fusion, the state of the art of each research field will be demonstrated according to the specific part of my research work. Note that although not all of these previous works are proved to be effective to heterogeneous information fusion, the reason why I include them all in this thesis is to give an general review of related researches and to let readers have a better point of view of why and how I conduct my research work, in other words: the motivation and methodology. To be specific, section 2.2 will focus on the previous work in the field of visual information acquisition, in which salient object detection and autonomous classification works are addressed. In section 2.3, previous detection models of salient sound will be reviewed and state of the art researches related to the environment sound recognition will be discussed. The section 2.4 will accounts on the autonomous fusion research based on the audio-visual fusion methodology and its application in the field of robotics. Furthermore, research works which related to the knowledge level fusion and artificial cognition that motivate my work are presented as well.

Inspired by the perception mechanism of human beings, a practical solution is to apply the saliency principle in heterogeneous feature extraction work and to form generalized saliency information of the environment in an audiovisual way. So far, as few related research has been successfully conducted to provide practical machine awareness ability for autonomous environment perception, saliency based information acquisition techniques are considered to be the potential possible approach. Hence, saliency principle will be introduced first as the fundamental basis in the following sections. The feature acquisition works of both audio and image signals will be illustrated individually, along with the state of the art research works which related to the audio-visual fusion based environment perception for machines.

The original definition of saliency is to describe an object that is prominent or conspicuous, which arises from the contrast between object and its neighborhood. The saliency issue was studied more frequently in visual system, such as human vision and computer vision, and less in auditory system as well, where the saliency principle is used as the artificial representation or bio-inspired modeling of saliency mechanism in human perception system, especially the "bottom-up" pattern (Frintrop et al., 2010).

Followed by the pioneering work done by (Koch and Ullman, 1985) and (Itti et al., 1998), many research works have been done in the field of saliency detection so far, including visual and auditory. Most of the existing works are based on different kinds of computational models which simulate the perception mechanism of human beings instead of purely bio-inspired methods in the field of visual saliency research, and several auditory mathematic representations have been developed following the same motivation of saliency based perception to mimic the human hearing system in sound signal processing. Therefore, I would like to give a general investigation in this chapter which focus on the salient information perception works that related to the subject of artificial awareness based autonomous environment perception. However, not only the previous researches about saliency based information acquisition but also the related works of audio-visual fusion and knowledge fusion will be illustrated in this chapter.

2.2. Autonomous Object Perception using Visual Saliency

2.2.1. Introduction

Visual stimulus is the most important information source for human perception system and perhaps the only solution for understanding the real world vividly and comprehensively. As the artifact that records the visual information, image acquired by man-made machines is the artificial representation of human visual perception system. In this thesis, the image refers to the two-dimensional digital signal that could be processed by image processing techniques on any ordinary computer platforms. Since the research of image processing has been well studied by researchers for many years and highly specialized developed according to different applications, it is impractical to investigate them all from every research field here. Nevertheless, image processing techniques that based on saliency principle and applied in robot perception scenarios will be demonstrated in general.

2.2.2. Object Detection using Visual Saliency

2.2.2.1. Basic Visual Saliency Model

Visual saliency (also known as visual attention in other literature (Liang and Yuen, 2013) refers to the perceptual quality that highlighted a region of image which stand out with respect to their neighborhood or unique in attributes relative to other regions, and to capture attention of observer (Itti, 2007) or applied for predicting interesting locations that human beings are likely to focus (Zhao and Koch, 2013). This has found to be efficient as first step in plenty of application scenarios, such as image segmentation ((Han et al., 2006), (Ko and Nam, 2006) and (Li et al., 2011b)), object recognition ((Rutishauser et al., 2004), (Walther et al., 2005) and (Gao et al., 2009)), image and video quality assessment ((Liu et al., 2013) and (Wang et al., 2004)), object detection ((Marfil et al., 2009), (Yanulevskaya et al., 2013) and (Yeh et al., 2014)), key frame extraction ((Lai and Yi, 2012) and (Ejaz et al., 2013)), to mention only a few.

The first computational model for saliency detection was proposed by (Koch and Ullman, 1985), in which biological inspired computation architecture is proposed. The motivation of this architecture is based on the feature integration theory (FIT) which presented by (Treisman and Gelade, 1980), and the output image is a two-dimensional topographical map which defined as the saliency map. As the key contribution of their model, the saliency map is yielded by computing the differences between basic features (i.e. color, intensity) of image from a scene and the most salient object for attention will be selected by using the Winner-Take-All (Lee et al., 1999) network based on an early representation mechanism. Inspired by this model, a more practical and easy implemented computational model was presented by (Itti et al., 1998), in which authors demonstrated that typical visual neurons are most sensitive in center-surround regions, and with this assumption, many research works have been conducted to implement and testify the efficiency of center-surround mechanism for saliency detection. For example, (Gao et al., 2007) and (Mahadevan and Vasconcelos, 2010) propose similar saliency detection methods that can be applied in videos and images based on the hypothesis that saliency is the result of optimal discrimination between center and surround stimuli at each location of the visual field. A set of visual features are collected from center and surround windows and the locations where the discrimination between the features of the two types can be performed with the smallest expected probability of error are declared as most salient. Background subtraction then reduces to ignoring the locations declared as nonsalient.

(Klein and Frintrop, 2011) introduce a computational approach for saliency detection based on the center-surround divergence, in which the saliency computation is divided into two steps: first, basic features of image are analyzed on different scales; secondly, the center-surround contrast distribution is determined by using the Kullback-Leibler Divergence (KLD) as Information Theory saliency feature. Thus, the conspicuity of a region in image can be rated by combine these two features together. Experiments show its good performance in processing typically saliency object database especially for small objects instead of the natural images.

However, the center-surround mechanism based saliency detection approaches are depending on the contrast of three basic features (i.e. intensity, color, and orientation) to emphasize the distinctiveness of certain region which reflects the local spatial-discontinuity, thus, this is called the local contrast (Yeh et al., 2014) since the center-surround saliency is calculated within a sliding window. The local contrast based approaches involved majority of saliency detection techniques, which different local features will be used or combined with features obtained from other scales to evaluate the existing probability of the salient object. For example, (Bruce and Tsotsos, 2009) proposed a computation framework built on a first principles information theoretic formulation dubbed Attention based on Information Maximization (AIM) for visual saliency, in which the Shannon self-information associated with each pixel location in the context is used to indicate the content of interest.

In the work of (Alexe et al., 2010) a Bayesian framework based on image cues for salient object detection is described, in which the image cues including multi-scale saliency, color contrast, edge density and superpixels straddling are integrated into a Bayesian classifier. This method is said to be useful in weakly supervised scenario, where object locations are unknown. Particularly, the concept of objectness is initially proposed to improve the accuracy and efficiency of saliency detection, in which the objects will be distinguished from the background. In (Ma and Zhang, 2003) a image attention analysis framework based on both “bottom-up” and “top-down” mechanisms is presented, in which a saliency map is derived by using local contrast analysis and the attended area is extracted from saliency map by using a fuzzy growing method as the simulation of human perception. Supported by user study results, this approach is proved to be computational efficient and effective in accuracy.

Despite all the successful works, argued by (Achanta et al., 2009) and (Cheng et al., 2011), the local contrast based methods share a minor limitation that the obtained saliency map could highlight the region near object boundary instead of the whole salient objects. Accordingly, the global contrast based methods are increasingly developed in which information features from all pixels of the entire image are considered and processed to yield a global saliency representation. In (Hou and Zhang, 2007) a spectral residual based image saliency detection method is proposed, in which the saliency map is yielded by using the log Fourier amplitude spectrum and the spectral residual of image is calculated. The method is in spectral scale so prior knowledge is not required and provides the ability of generality to cover unknown features. Experimental results show that the saliency map covers all the salient objects to static natural images. A similar approach refers to (Guo et al., 2008), in which the phase spectrum of the Fourier transform is proposed for detecting the location of salient area instead of the amplitude spectrum. Consequently, authors extend the phase Fourier transform (PFT) to the phase Quaternion Fourier Transform (QFT) which consider the global spatiotemporal characteristic of the image and video as well, therefore, the method is independent of parameters and prior knowledge and provide good robustness to white-colored noise which supported by experimental results.

(Zhang et al., 2008) proposed a saliency using natural statistics (SUN) model using Bayesian framework. In this framework, the bottom-up saliency was measured by means of the inference of generalized Gaussian distributions over the entire image based on the natural statistic which obtained in advance from a collection of natural images. This overall visual saliency is the pointwise mutual information between the observed visual features and the presence of an object. The method shows that the natural statistics, as the global contrast feature, is very crucial to saliency computation when the model is learned over times. Note that, in the work of (Achanta et al., 2009) a frequency-tuned salient region detection approach is introduced as the improved version of the authors' previews work in (Achanta et al., 2008). By analyzing the property of spatial frequency content from the original image, appropriate range of spatial frequency is retained by using the difference of Gaussians (DoG) approach and combined with the low level image feature (i.e. color and luminance) to compute the final saliency map.

In robotic research field, image saliency detection is also widely used in different application scenarios as an efficient tool for robot vision processing. For instance, a simple visual saliency model for robot camera is described in (Butko et al., 2008), where the fast and computationally-lightweight adaptation of traditional Bayesian approach is demonstrated. Experiments in real world show that, this model is able to provide saliency maps in real time on a low-end computer and matches human eye fixation data well, thus makes the proposed method promising for social robotics applications. In (Chang et al., 2010) a robotic vision navigation and localization system is deployed based on the bio-inspired image features of salient region and gist, the robot is able to execute user's command correctly and go to a goal location after trained for just one time. As the salient regions from images which collected by camera are the evidences for landmark exists, a stand-alone localization process is performed by robot and the experimental results show the proposed method is not sensitive to the temporal order of input images. Recently, (Scharfenberger et al., 2013) proposed an object detection algorithm based on the saliency histogram features of image for robot vision. The histogram of saliency values is used to remove the geometrical information from saliency maps and, by applying Principal Component Analysis (PCA) approach, yield the existence of the interesting object. Though the performance is claimed to be outperformed than other state of the art approaches by authors, experimental result images with only single object are provided. In (Ramik et al., 2013) and (Ramik et al., 2013), the perceptual saliency detection approach is realized to simulate the artificial perceptual curiosity and to enable unsupervised extraction as well as subsequent learning of a previously unknown object by the machine in a way that realizes perceptual curiosity. The technique is fully autonomous; it allows the robot to perceive the surrounding environment and learning new knowledge by interaction with human tutor, however, the sound information of the objects are not used in this research.

Though the previous mentioned researches have been proved to be effective in detecting visual saliency, it can be obviously seen that both local and global contrast based methods partly represent the unified saliency property of an image in certain levels or scales. Consequently, the optimal solution should take both local contrast feature and global contrast information into consideration and integrate them with other saliency features to form the saliency map.

Followed by this motivation, (Wang et al., 2013b) propose a multi-spectrum based saliency detection algorithm in which the near-infrared clues are incorporated into the framework and the color contrast is combined with texture saliency feature to export the saliency map simultaneously, the performance of this algorithm is proved to be promising by testified on natural images. Meanwhile, a bottom-up stimulus driven motivated saliency region detection method by using joint spatial-color constraint and multi-scale segmentation technique is presented in (Xu et al., 2013). Three components are used to estimate the pixel-level saliency for the input image which are spatial constraint for global contrast of “center-surround”, intensive contrast of red-green and blue-yellow double opponency and similarity distribution of a pixel for object and background distinction, respectively. Though the method can uniformly highlight salient regions with full resolution and suppress the surrounding background even in the large object case, the performance of which will significantly degrade if the foreground is similar to background. (Yeh et al., 2014) described a novel salient object detection approach based on region growing and competition process by propagation the influence of foreground and background seed-patches, which is the optimal combination of local distinctive regions and global homogeneous parts. Experimental results from both well-known benchmark datasets and natural images show that significant saliency results could be obtained by the proposed approach and the shortcomings exist in the work of (Xu et al., 2013) could be overcome. In (Harel et al., 2007) a graph based visual saliency (GBVS) model is proposed, in which each node represents a lattice and the connection between two nodes is proportional to their dissimilarity where the contrast was inferred by Markov chain. Furthermore, (Goferman et al., 2010) combined local contrast, global contrast, visual organizational rules, and high-level cues to form a new type of saliency called context-aware saliency. (Li et al., 2011a) proposed to combine global contrast to frequency domain and local contrast to spatial domain for the generation of a saliency map.

2.2.2.2. State of the Art Methods

Despite that both local and global contrast could be merged with other saliency features to achieve better saliency detection accuracy, the previous mentioned works are basically conducted based on the center-surround mechanism which inspired by the bottom-up pattern of visual perception. However, from the perspective of human visual awareness, it is obvious that we always intend to focus on the most informative region or object in an image in order to efficiently analyze what we have observed (Yarbus, 1967). Since the salient objects extracted by using previous mentioned visual saliency models are derived from bottom-up based center-surround operation, most of the proposed saliency models are the results of mathematically computation and not in the optimal representation forms, which means that the information redundancy is included. Motivated by the theory of information which proposed in (Shannon, 1948), (Attneave, 1954) suggested that the statistical properties of images are correlated with certain aspects of visual perception, which means that the ultimate goal of visual perception is to produce an efficient representation of what human observed from the surrounding environment. (Barlow, 1961) argued that the informational coding efficiency is an important constraint on neural processing since neurons encode the information in a very efficient way in order to effectively utilize the available computing resources. In the work of (Barlow, 1972) observation has been obtained that active neurons at later stages of processing are less than the earlier stages, which inferred that the coding strategy of information in the nervous system has higher degrees of specificity. The coding strategy is further known as the redundancy reduction principle, which leads to the notion of sparseness (Field, 1987).

Biological evidences from (Hubel and Wiesel, 1968), (De Valois et al., 1982), (Jones and Palmer, 1987) and (Parker and Hawken, 1988) have demonstrated that, the receptive fields of simple cells in the primary visual cortex (V1) of mammalian could be characterized as being spatially localized, oriented and band-pass. The works of (Olshausen, 1996) and (Olshausen and Field, 1997) have shown that such response properties could be similarly represented by applying the sparse coding strategy on natural images without imposing a particular functional form on the receptive fields. They built the model of images based on the overcomplete basis functions and adapted them in order to maximize the sparsity of the representation. The set of functions which generated after training on tremendous image examples strongly resembled the spatial receptive field properties of simple cells, which means that neurons in V1 are configured to represent natural scenes in terms of a sparse code ((Olshausen and Field, 2000) and (Simoncelli and Olshausen, 2001)). Many other research efforts including (Vinje and Gallant, 2000), (Olshausen, 2003), (Hoyer and Hyvärinen, 2002) and (Hoyer, 2003) have also been made to show that the sparse representation better described the perceptual response of human vision perception system, while the effectiveness and efficiency of sparse coding in the application of image processing has also be studied in (Olshausen and Field, 2004), (Lee et al., 2006) and (Elad and Aharon, 2006a).

In (Bruce and Tsotsos, 2005), they applied information maximization principle to the image in order to locate the salient object, in which the sparse representation of high-level features which performed by the independent component analysis (ICA) is presented to facilitate the selection. Since the coefficients of ICA basis functions are statistically independent and non-Gaussian, the image is represented by the ICA basis functions in the fashion of sparse which similar to the characteristic of human vision perception. Motivated by Bruce's work, (Sun et al., 2010) proposed an image saliency detection approach based on the short-term statistics of single images instead of using the large scale natural statistics, in which saliency property is measured by the feature activation rate (FAR). Experiments show that the proposed method could yield a more accurate performance than traditional method due to the strong biological plausibility.

Similarly, many works have been conducted by combing the ICA basis functions which learned on a large number of randomly selected image patches with other features or based on different algorithms to measure the visual saliency property. For example, (Wang et al., 2010b) defined the site entropy rate (SER) of the random walk on the graph structure to measure the visual saliency property based on the principle of information maximization, in which the sparse coding bases are learned by using ICA algorithm. Though experiment results have demonstrated a better performance in receiver operator characteristic (ROC) curves and the areas under ROC (AUC) than traditional methods on both images and videos, this method is simply modeled based on the dissimilarity of spatial and temporal features from images. In (He et al., 2011) three attention-oriented features derived from sparse coding are combined for visual saliency detection, in which the sparse codes are learned from eye-fixation patches by using ICA algorithm and all the features are extracted based on the sparse coding coefficients. The experiment results have shown good detection performance, the expected objects are however distinctive as the background regions are not complex or contain interference objects.

Recently, (Yang et al., 2013) proposed to fuse local saliency feature and global saliency feature from each color channel together based on the sparse representation of image for saliency detection, in which the local saliency feature is calculated as the average dissimilarity between a center patch and its rectangular surrounding patches and the global saliency feature is obtained simply by using the dissimilarity between a local patch and the center patch. The saliency property defined in this work is simply based on the distance between a given patch and the center patch from different scales as well as different color component channels. Therefore, the contrast of center patch and other patches are the crucial factor that could influence the detection performance which limits its application in real world. It can be observed from the experimental results that the ground truth salient objects are generally located in the center of the images with distinctive color representations. Accordingly, the research conducted in (Anwar et al., 2014) could be seen as a similar work from the perspective of center surround difference (CSD) mechanism, in which the visual saliency detection is based on the normalization of multi-scale saliency features. To be specific, they proposed to use the adaptive sparse representation to boost the performance of the CSD operation and the nonlinearly integrated of color and spatial representation is presented to better capture the structure of underlying data. However, the saliency detection performance could be decreased when multiple objects are existed in the image and the background is complex or similar to the salient objects.

The comprehensive research of (Han et al., 2013a) is considered to be similar to my work. They proposed a probabilistic framework for visual saliency detection using sparse coding representation, in which joint posterior probability of pixel rarity and objectness is presented to achieve the saliency value. The rarity probability of a pixel is obtained by using the global contrast which derived from the sparse coding based feature vector with respect to the Bayesian rule, while the objectness probability of a pixel determines whether the pixel belongs to an object or not. The characteristic of objectness of the salient object is modeled in three aspects to represent the inherent property of an object which are compactness, continuity and center bias, respectively. Since the object will have a group of pixels in image, the Gaussian mixture models (GMMs) is applied to model the responses of different filters. The obtained saliency maps have shown distinctive results with salient objects being properly extracted, while the performance of proposed approach is evaluated with respect to ROC curves and AUC factor which outperforms other algorithms. However, the average time cost is not optimal because of the usage of GMMs, which leads to the further improvement on computational complexity.

2.2.3. Autonomous Object Recognition and Classification

2.2.3.1. Image Feature Acquisition

In various computer vision applications, to automatically recognize and classify the objects, the fundamental work is to locate the expected objects in given images. Though quantities of features have been proposed in the field of image processing to describe the interesting or unique information behind each pixel, such as the feature of edge, corner, blob and ridge, many other image features have been developed to better represent the object.

From the recognition as well as classification point of view, the state of the art image features of object that frequently used nowadays can be illustrated in two general categories which are local point features and statistic features. In this section, several well-developed feature extraction techniques will be presented to give a brief introduction of current image feature acquisition work which is the foundation of the object recognition research.

2.2.3.1.1. Local point feature

The local point features are obviously localized features that represent the local characteristic of an image patch, such as interest point, corner point, local orientations or gradient and local pattern of pixels. Compared to the corner points or edges, the majority of popular used local point features in object recognition scenario are more complex and designed to be translation, scale and rotation invariant, some of which could also be partly invariant to illumination and robust to local geometric distortion, such as the scale invariant feature transform (SIFT) based feature, the speed-up robust feature (SURF) and the local binary pattern (LBP) feature.

1) SIFT feature

The SIFT feature is proposed for image-based recognition application based on the scale invariant feature transform which is initially developed by (Lowe, 1999) and further improved in (Lowe, 2004). This feature is considered as a local point feature because it is performed on a local image patch to obtain the description of local patch. In particular, the SIFT feature uses different of Gaussian (DoG) to determine the scale and location of feature while a pyramid structure is applied to simulate the multi-scale characteristic of image.

2) SURF feature

Inspired by the SIFT feature, the SURF feature is proposed by (Bay et al., 2006) to simulate the receptive-field-like response in the neighborhood of an interest point. To construct the pyramid image, the Hesssian matrix of each pixel is used in SURF instead of DoG scale space. However, according to the calculation of SURF and SIFT feature, the differences of them can be summarized in threefold which are shown in the Table 1 as follows:

	SIFT	**SURF**
Key Point Detection	Multi-scale images convoluted with DoG	Integral image convoluted with multi-scale box filter
Orientation	Calculated by histogram of gradient within rectangle neighborhood of key point	Calculated by the response of Haar wavelets in x, y directions within circular neighborhood
Feature Descriptor	Divide 20*20 pixel region into 4*4 sub-region, calculate 8-bin histogram in each sub-region	Divide 20*20 sigma region into 4*4 sub-region, calculate Haar wavelets response of 5*5 samples

Table 1: The differences between SIFT and SURF.

3) LBP feature

To better represent the texture feature for object recognition, the LBP feature

(Ojala et al., 2002) is proposed as a particular example of what have been described in the Texture Spectrum model which was proposed in (He and Wang, 1990). As a nonparametric approach, the LBP feature describes the local texture characteristic of an image by comparing each pixel with its neighborhood pixels to form a LBP codes.

The original LBP descriptor is defined within a 3*3 pixel window that the value of 8 neighborhood pixels are compared with the value of center pixel by subtracting the center pixel, the results with positive values are encoded with 1 while others with 0. Therefore, an 8-bit binary code could be obtained by concatenating the binary code of each pixel in a clockwise direction which starts from the top-left neighborhood, the decimal value of which is then used to represent the texture information of the pixel, thus the histogram of LBP labels calculated over a region could be used as the texture feature of an image. The basic LBP operator is shown in Figure 1 to illustrate the mentioned process.

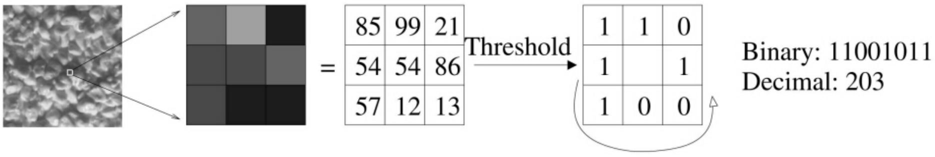


Figure 2: The basic LBP operator. Adopted from (**Ahonen et al., 2006**)

However, the limitations of the basic LBP operator are obvious. The 3*3 window could only cover a small area which makes it impossible to capture proper features for large-scale structures and different frequencies of texture pattern. In order to improve the robustness of basic LBP operator in dealing with textures at different scales and increase the calculation simplicity, two major extension have been proposed which are LBP(*P,R*) and the uniform pattern LBP (Ojala et al., 2002). The LBP(*P,R*) are defined as a set of sampling points evenly spaced on a circle around the center pixel to better represent the neighborhoods of different sizes. In particular, the notation (*P,R*) denotes a pixel neighborhood of *P* sampling points on a circle of radius of *R*, in which bilinear interpolation is applied when sampling point does not fall within the pixels. In Fig. 2, an example of the circular (8,2) neighborhood is shown to demonstrate the bilinear interpolation.

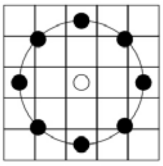


Figure 3: The circular (8,2) neighborhood. Adopted from (**Ahonen et al., 2004**)

While in the circular formation of LBP(*P,R*), the LBP is defined to be uniform if the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa. For example, the patterns 00000000 (0 transitions), 00001111 (1 transitions) and 11100111 (2 transitions) are uniform, while the patterns 11011011 (4 transitions) and 10100101 (6 transitions) are not. As observed in (Ojala et al., 2002), uniform patterns accounts for a bit less than 90% of all patterns in the (8,1) neighborhood and for 70% in the (16,2) neighborhood. Consequently, the number of original LBP operator could be decreased from 2^P labels to $P*(P-1)+2$ labels by accumulating the non-uniform patterns into one pattern. Thus, the notation of uniform pattern LBP operator can be defined as the $LBP_{P,R}^{u2}$, in which the subscript denotes the (*P,R*) neighborhood while the superscript *u2* stands for using only uniform patterns. The LBP features are frequently used in many scenarios because it invariant to contrast and illumination, such as face detection and recognition (Ahonen et al., 2004), texture classification (Fernández et al., 2013) and object recognition (Satpathy et al., 2014). Note that, as the main shortcoming is that it is sensitive to noise and variations of small pixel values, proper improvements have been made based the combination with other features.

2.2.3.1.2. Statistical Feature

Different from the local point feature which based on the detection of the interest local pixels, the statistical feature of image focus on the statistic distribution that can represent certain properties of objects. Although lots of statistical feature from image domain have been proposed to describe the uniqueness of object as well as proved to be effective in their experiments, only the one that generally confirmed to be effective in the majority of applications is considered to be worth of demonstration. Therefore, the GIST feature and histogram of oriented gradient (HOG) feature are demonstrated as the typical representation of statistical feature of an image.

1) GIST feature

The GIST feature was proposed in (Oliva and Torralba, 2001), in which the term of GIST was used as an abstract representation of the scene that could spontaneously activates memory representations of scene categories which was initially discussed in the work of (Friedman, 1979).

Inspired by the human perception characteristic, the GIST feature is designed to be the low dimensional representation, or rather the spatial envelope, of the real world scenes without segmentation of image and processing of individual objects or regions. In order to characterize several important statistics about a scene, a set of perceptual dimensions including naturalness, openness, roughness, expansion and ruggedness are proposed to represent the dominant spatial structure of a scene. The GIST feature is computed by dividing the image into a 4*4 grid to obtain the orientation histograms. Specifically, the image is convolved with an oriented filter at different orientations and scales in order to obtain the high- and low-frequency repetitive gradient directions and yield the histogram of each orientation and scale as the feature. As a global image representation, the GIST feature has been used with other features in the application of image search (Kovashka et al., 2012), object recognition (Zhao et al., 2011), context classification (Farinella et al., 2014) and image segmentation (Zhou and Liu, 2014).

2) HOG feature

In (Dalal and Triggs, 2005), the HOG feature was firstly proposed for pedestrian detection in static images. The general concept of HOG feature is that the appearance and shape of local object can be described by the distribution of intensity gradients or edges, which is in fact, the statistical information of gradients that mostly exist in the edges. Different from other descriptors, the HOG feature is obtained by accumulating a local 1-D histogram of gradient directions and edge orientations over the pixels that within a dense grid of uniformly spaced cell. For better invariance to illumination and shadowing, the local histograms of cells are contrast-normalized within larger spatial overlapped blocks in the detection window.

Since the HOG feature is calculated based on the localized cells, it is invariant to the geometric and photometric transformations of images while except for the object orientation. Meanwhile, as suggested by the work of (Dalal and Triggs, 2005), coarse spatial sampling, fine orientation sampling and strong local photometric normalization could allow small body movement of pedestrians without contaminating the detection performance as long as they maintain a roughly upright position. Therefore, the HOG feature can be used in the detection and recognition of objects with small variations in shape and illumination, and is particularly suited for human body detection in images and videos (see (Dalal et al., 2006) as a classical research). Thus, it is further used in the detection and recognition of particular part of human including the head-shoulder detection (Zeng and Ma, 2010), hand gesture recognition (Feng and Yuan, 2013) and face recognition (Li et al., 2013).

2.2.3.2. Current Recognition Models

Regarding the recognition of object, most of the state of the art works that have been conducted are based on the recognition models that previous learned on a dataset or large-scale training examples to accomplish the further recognition task. However, according to the differentials in learning method of different models, the majority of recognition models could be generally divided into static model and dynamic model. Suggested by the terms, the main difference between static model and dynamic model is that whether if the model can be update after the training process. To be specific, the static model is learned in the training process with labeled data without any update, while the dynamic model can update itself by certain learning approaches with new data or even unlabeled data.

Though there are many recognition models have been presented in literatures and proved to be effective, each kind of model has its own advantages and limitations as well. From static model point of view, the bag-of-word (BoW) (also known as bag-of-visual-word, bag-of-feature and bag-of-visual-feature in other literatures) model is the well-researched one among other models. The BoW model is originally motivated by the texture recognition issue, in which the image representation of particular texture can be characterized by the histogram of image textons as the feature vector of texture class (Cula and Dana, 2001). Thereafter, the BoW model was applied in the field of object recognition by (Sivic and Zisserman, 2003) and (Sivic and Zisserman, 2006) in the term of 'bag-of-visual-words' (BoVW) and has exhibited good performance.

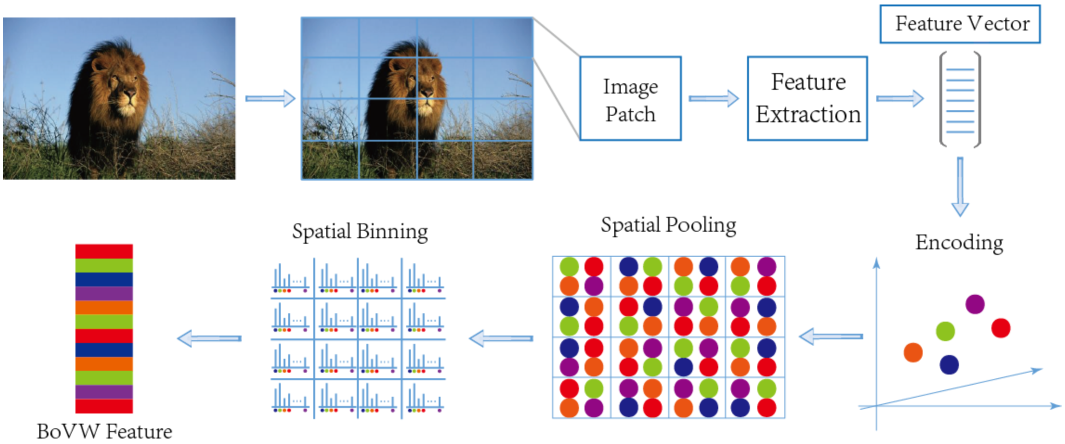


Figure 4: The flow diagram of classic BoVW model. Figure inspired by (**Chatfield et al., 2011**)

The underlying assumption of the BoVW model is that, the image of an object can be represented as an unordered collection of image features that derived from all the local image patches, in which a 'vocabulary' or 'bag' of 'visual words' is generated for specific image set of particular object by using the clustering method. Therefore, histograms can be achieved for objects by projecting the corresponding visual features onto the obtained 'bag of visual words' or code words as the BoVW feature for further object classification and recognition process. The flow diagram including several key components and processes of the BoVW model is graphically demonstrated in Figure 3.

Recently, the BoW/BoVW model is a popular method in several applications including scene recognition (Cakir et al., 2011), image classification (Sujatha et al., 2012) and object detection (Lampert et al., 2009), due to its simplicity and efficiency. Meanwhile, many adaptation works have been presented based on the classic BoW model to improve the performance of model. For example, (Liu et al., 2014a) proposed a hierarchical BoVW framework for image classification, in which the coding vectors obtained in the first layer of this framework are treated as higher level features and a soft voting based coding strategy is applied. The proposed framework is flexible to be combined with other coding and pooling approaches for particular tasks because of the simplicity in structure. Compared to the traditional BoW model, the experimental results of this framework have demonstrated that the accuracy of classification can be improved especially when there are fewer code words in the first layer. In (Ji et al., 2013) an object-enhanced feature generation mechanism based on the BoVW model for scene image classification is proposed, in which the local feature of salient object region has higher weight than the background in order to enhance the importance of non-background region without any label operation. The enhanced features of all the object regions are merged with local feature of background regions to form the final feature vector in a traditional BoW model manner. The experiment results conducted on popular scene datasets have shown a 1.0-2.5% increase compared to the original feature, which demonstrated the effectiveness of combining visual saliency with the BoW model.

Moreover, (Nanni and Lumini, 2013) combined the random selection of interest regions, the heterogeneous set of texture descriptors and the BoW model of which different codebooks were generated by several *k*-means clustering runs, together to accomplish the object recognition task. Different from other approaches, the proposed system uses different texture descriptors as the local descriptors to obtain the local features from the overlapping sub-windows of image, while the global features are calculated within five specifically selected regions from the original image. In their work, the representation redundancy and robustness of features have been taken into account by using the ideas from PCA-SIFT to reduce the feature dimension as well as by using the combination of different codebooks and texture descriptors. The results of experiments which conducted on four different widely used datasets have shown a consistent improvement of the stand alone approaches and remarkable results could be achieved with respect to other state of the art approaches, which demonstrated the advantage and efficiency of using the idea of feature dimension reduction. (Huang et al., 2014) proposed a multiple spatial pooling method to capture the global spatial structure for object recognition based on the BoW model, in which the pooling matrix was constructed by employing multiple Gaussian distributions in order to alleviate the sensitivity to object shifting. The validation results on popular dataset have shown a best increase of 22.8% in recognition performance than other related approaches while computational cost is acceptable, which argued that the global spatial structure of the object is crucial to the recognition task.

2.3. Salient Environment Sound Perception

According to the difference of property between sound sources, sound signals can be generally categorized into several kinds, such as human speech, musical sound, vibration signal and environmental sound. Thus, sound signal processing plays an important role in the research field of voice recognition, sound event analysis, music recognition, fault diagnosis and environment perception. Compare to image which generated by visual perception system, sound signal is actually a mechanical wave that needs medium to be transmitted through. Therefore, sound could be seen as a heterogeneous signal that should be processed differently according to image signals. Though there are various kinds of sounds exist in modern human society, only those sounds that related to environmental objects are essential for environment perception of machines. Therefore, the detection and recognition of environmental sounds are researched in this thesis.

2.3.1. General Introduction

During the exploration of real world in everyday life of human beings, the perception process is carried out in a very effective and efficient way due to the existence of saliency-driven selective attention mechanism. Despite the human visual perception is dominant in most scenarios, auditory perception system could provide information of specific events or objects beyond visual range and bypass the obstacles. Hence, sound signal will enable mankind to be aware of and avoid danger beforehand or when human vision is not available in certain environment. However, modern artificial machines like industrial robot still could not perceive its surrounding environment as intelligent as human does, because the perception process is mostly based on the visual sensory information rather than auditory.

Another reason is that environmental sound signals are vary in both spatial and frequency properties and requires complicated techniques for processing. Therefore, it is essential to provide saliency-driven approach for machine to form the artificial awareness ability with respect to the cost of both time and computational resources.

2.3.2. Auditory Saliency Detection

2.3.2.1. Motivation

Currently, research works of saliency-driven auditory perception are mainly based on the auditory saliency map first proposed in the pioneering work of (Itti et al., 1998). In Itti's auditory saliency map, the salient sound signal will be represented by the visually specific region with respect to its time-frequency characteristic. As argued by (Li et al., 2008) that images and sounds have aesthetic connections in human perception system, the auditory saliency can be reformed for perceiving the bottom-up saliency mechanism in the visual domain.

Moreover, sound signal could provide omnidirectional information about specific event or objects happened or existed in surrounding area, as well as ignoring the influence of obstacles and make the perception more vividly. Saliency-driven auditory perception is also an important research issue regarding the realistic demand in the field of machine awareness. For instance, when a robot is exposed to certain emergency situation like explosion, many salient external stimuli from the same event or object will be received synchronously including the image and sound of either explosion or alarm triggered by the explosion. Since image signal could be blocked by other objects, the sound signal can play a vital role in detection of salient event.

2.3.2.2. Saliency Detection Models

From auditory saliency detection point of view, inspired by the research work of visual saliency map, (Kayser et al., 2005) initially proposed an auditory saliency map for salient sound detection, in which the auditory saliency model is based on the spectrogram of the input sound signal and the spectrum based auditory saliency maps are obtained by using center-surround difference operator (see (Itti and Koch, 2001)) to transform auditory saliency into image saliency for further analyzing. Though experiment results have shown that this model is able to find a salient natural sound among background noise, only the visual saliency features from the image of sound are considered while the information of auditory signal have not been taken into account. The second model is proposed by (Kalinli and Narayanan, 2007) as an improvement of the Kayser's work, in which two more features of orientation and pitch are included and a biologically inspired nonlinear local normalization algorithm is used for multi-scale feature integration. Its performance is tested in sound signals of read speech and is proved to be more accurate with 75.9% and 78.1% accuracy on detecting prominent syllables and words. However, in Kalinli's work, the sound sources are selected from the broadcast read speech database and no environment sound tracks from real world have been used. The third model is proposed by (Duangudom and Anderson, 2007), in which the spectro-temporal receptive field models and adaptive inhibition are applied to form the saliency map. Supported by the results of the experimental validations consisting of simple examples, good prediction performance can be obtained but the model is still not verified on real environmental sound data which limits its application in industrial manufacture.

Recently, (Kim et al., 2014) considered the Bark-frequency loudness based optimal filtering for auditory salience detection and researched on the collecting annotations of salience in auditory data, in which linear discrimination was used. Though the experiment results shown 68.0% accuracy, the sound signals for validation are collected from meeting room recordings. This means that only indoor environment is considered. (Kaya and Elhilali, 2012) proposed a temporal saliency map approach for salient sound detection based on five simple features, in which the saliency features only from time domain are considered. Though the test results have shown that this method is superior to Kayser's work, the lack of frequency contrast and other auditory information limit its application. (Dennis et al., 2013b) proposed a salient sound detection approach based on the keypoints of local spectrogram feature for further sound event recognition in which the overlapping effect and noise conditions are considered. Though the experiment results have shown a clear detection output on multiple sound sources, only simple sound examples are used for experimental test and real environment sound signals are not included.

2.3.3. Environmental Audio Information Perception

2.3.3.1. Feature Extraction of Audio Signal

So far, many research works have been done in developing approaches that could automatically analyze and recognize the sound signal. Currently, all the works are mostly based on the characteristics, or rather the features, of sound. The features of sound that used in early works are often referring to the descriptors that can be calculated by using sound data from time or frequency domain, such as root mean square, zero crossing rate, band energy ratio and spectral centroid (Widmer et al., 2005). However, although these kinds of features can be easily derived from sound data, they are still lower-level features that need sophisticated approaches to process and could not represent the human acoustic characteristic for higher-level fusion work. According to (Cowling and Sitte, 2003) the sound signal from natural world can be either stationary or non-stationary, a more general way of categorizing sound features proposed by (Chachada and Kuo, 2013) is to divide them into two aspects according to the differences of the extraction methods, which are as follow:

1) Stationary features

Stationary features including both temporal and spectral features, such as the Zero-Crossing Rate (ZCR), Short-Time Energy (STE), Sub-band Energy Ratio and Spectral Flux (Mitrovic et al., 2010), which are frequently used because they are easy to compute and could be concatenated with other features. Meanwhile, as the computational resemblance of human auditory perception system, the Mel Frequency Cepstrum Coefficient (MFCC) is often used in most of the human-voice related audio signal processing scenarios like speech and music recognition. Other widely used cepstral features including Linear Predictive Cepstral Coding (LPCC), Homomorphic Cepstral Coefficients (HCC), Bark-Frequency Cepstral Coefficients (BFCC) and the first derivative of MFCC which is ΔMFCC as well as the second derivative of MFCC (ΔΔMFCC).

2) Non-stationary features

In most of the research works, non-stationary features are referring to: a) Time-frequency features derived from time-frequency domain of audio signal, which are the spectrograms generated by Short-time Fourier transform (STFT) as well as the scalogram generated by Discrete Wavelet Transform (DWT) or Continuous Wavelet

Transform (CWT). b) Sparse representation based features which are extracted by the Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP) approach based on the atoms in an over-complete dictionary that consists of a wide variety of signal basis ((Chu et al., 2008), (Chu et al., 2009) and (Zubair et al., 2013)). c) Other time-frequency features, such as pitch range (Uzkent et al., 2012).

Recently, many works like (Mukherjee et al., 2013), (Chia Ai et al., 2012), (Chauhan et al., 2013) and (Samal et al., 2014) have been conducted in extracting the sound feature by using the computational model for further recognition processing, in which the MFCC and LPCC are frequently used to extract the human acoustic features from sound signal. In (Nirjon et al., 2013) a novel approach combined sparse Fast Fourier Transform which is proposed by (Hassanieh et al., 2012) with MFCC feature is proposed for extracting a highly accurate and sparse acoustic feature which could be used on mobile devices, the capability of this approach is proved to be efficient by the experimental results and could be expected to run in real-time at a high sampling rate. In the research work of (Murty and Yegnanarayana, 2006) a speaker recognition method based on the fusion of residual phase information and MFCC by using the auto associative neural network (AANN) (described in (Yegnanarayana and Kishore, 2002)) model is demonstrated and tested to be helpful in improving the performance of conventional speaker recognition system, experimental results show that the equal error rate of individually MFCC and residual phase is improved significantly. Meanwhile, similar research works are also carried out in (Nakagawa et al., 2012), in which the phase information is extracted by normalizing the change variation in the phase according to the frame position of input speech and combined with MFCC to accomplish the goal of speaker identification and verification. Moreover, (Hossan et al., 2010) propose an improved Discrete Cosine Transform (DCT) and MFCC fusion based acoustic feature with a Gaussian Mixture Model (GMM) classifier for speaker verification, experimental results show that the identification accuracy is better even with a lower number of features and the computational time is reduced as well. Furthermore, a saliency-based audio features fusion approach for audio event detection and summarization is presented in (Zlatintsi et al., 2012), where the local audio features are fused by linear, adaptive and nonlinear fusion schemes to create a unified saliency curve for audio event detection. The events combined with audio saliency features used for audio summarization, however, are manually segmented.

Nevertheless, despite all the research works in human voice related sound signal processing by using the acoustic feature of MFCC and other information proved to be successful, the research work of (Allen, 1994) shows that there is little biological evidence for frame-based features like MFCC, and that the human auditory system may be based on the partial recognition of features that are local and uncoupled across frequency, which means the local spectrogram features could be used to simulate the human recognition process. In other words, the spectrogram based spectrum image can be seen as the visual representation of sound signal and thus certain prominent image region could be processed as the image feature of sound for audio signal classification and recognition work.

Followed by this inspiration, the musical sound recognition researches have been put out in (Yu and Slotine, 2008), (Ghosal et al., 2012) and (Yu and Slotine, 2009), in which the sound features for classification and recognition are derived from the spectrograms which represents the time-frequency characteristics of different musical instruments as the texture images. The experimental results show that, performances of the proposed spectrogram-based classification algorithms are surprisingly good. In (Dennis et al., 2011) a spectrogram based information acquisition method is initially proposed for sound recognition and the robustness of recognition in mismatched conditions is improved, while in the research of (Dennis et al., 2013a) a more complete model of spectrogram feature based sound event recognition method is proposed. The general idea of this method is to use the keypoints that detected in the spectrogram and combined with the Generalized Hough Transform (GHT) (Ballard, 1981) for further recognition work. In (Kobayashi and Ye, 2014), an improvement version of local binary pattern (LBP) (proposed in (Pietikäinen et al., 2011)) based feature is presented. In this approach, to improve the robustness of LBP-based features to fluctuations in pixel values from the spectrogram image, the local statistics of pixel values including the mean and standard deviation are incorporated with the classic LBP-based features to form the feature vectors by using L2-Hellinger normalization. The classification performance of derived feature is tested and compared with other methods based on the real environmental sounds. The results show a 98.62% accuracy as well as the robustness to noise and low computation time.

2.3.3.2. Environmental Sound Recognition

Similar to the musical signals, the audio signals collected from environment can be also seen as the combination of foreground and background sounds (Raj et al., 2007). In most of the perception scenarios, foreground sounds are composed of signals that related to the specific acoustic events or objects while background sounds are more commonly generated by the acoustic scene. Therefore, the research orientations of non-speech environmental sound recognition can be divided into two categories: a) event based sounds, b) scene based sounds.

Though the research of sound scene recognition (see (Su et al., 2011) as an example) is important for understanding the environment, the scene based sounds can only be used to recognize the scene and are not able to provide comprehensive information for artificial awareness to achieve the perception task. Therefore, the recognition of event based sounds is more suitable for perceiving the information of what is happening in the environment and makes the artificial awareness possible.

Current research works of environmental sound recognition can be found in many application scenarios, such as audio retrieval ((Lallemand et al., 2012) and (Mondal et al., 2012)), robot navigation (Chu et al., 2006), surveillance (Sitte and Willets, 2007), home automation (Wang et al., 2008) and machine awareness (Wang et al., 2013a). Though the above mentioned works have proved that environmental sound recognition is essential and crucial to human life as well as the artificial awareness, most of the approaches are still focus on the theoretical level recognition works which based on the combination of previous mentioned audio features.

Recently, similar works like (Souli and Lachiri, 2011), (Souli and Lachiri, 2012a) and (Souli and Lachiri, 2012b), applied the non-linear visual features which extracted by log-Gabor filtered spectrogram and support vector machines approach to recognize the environmental sound. Note that, though experimental results show good performance of the proposed methods, the advantages and shortcomings between the log-Gabor filtered spectrogram and raw spectrogram are still need to be researched. Another spectrum image based sound signal processing application is for sound-event recognition which are demonstrated in (Kalinli et al., 2009) and (Janvier et al., 2012), both of which are motivated by the saliency based human auditory attention mechanism and the fusion of auditory image representation in spectro-temporal domain and other low level features of sound is applied. The experimental test data are collected from the actual audio events which occurring in real world and results show better performances of the proposed approaches with less computational resources. In (Lin et al., 2012) a saliency-maximized spectrogram based acoustic event detection approach is proposed, in which the spectrogram of audio is used as a visual representation that enables audio event detection can be applied into visually salient patterns, in which the visualization is implemented by a function that transform the original mixed spectrogram to maximize the mutual information between the label sequence of target events and the estimated visual saliency of spectrogram features. Although results of experiments which involved human subjects indicate that the proposed method is outperforming than others, since computational speed is prior considered in this method, the research of visual saliency features automatically extraction of audio events is not further conducted which lead to a limitation in robotic application.

Specifically, (Khunarsal et al., 2013) proposed a classification algorithm for very short time environmental sound processing based on spectrogram pattern matching. The local features of original environmental sound in spectrogram are derived after the signal is partitioned into several sub-signals by using a filtering window and concatenated as an input vector for classification, in which the *k*-nearest neighbor and neural network classifiers are deployed. The proposed method is proved to be effective with accuracy over 85% by experimental results, further, the performance of combinations with other auditory features (i.e. MFCC, LPC, and MP) are investigated as well and gain a maximum accuracy of 94.98%.

However, due to the unstructured property in most of the environmental sounds, it is obvious that the recognition accuracy of using spectrogram based image features could be sharply decreased especially when the background noise is relative strong, even if it combined with other spectral or temporal features. Therefore, many research works have been conducted based on the combination of acoustic features without using the spectrogram based image features. (Zhang and Li, 2013) proposed an ecological environmental sound classification method, in which a double-level energy detection (DED) based feature is presented. The feature is derived from the detection of the combination of two sub-detectors which perform frequency-domain energy detection (FED) and time-domain energy detection (TED), and merged with MFCC feature to yield the feature vector. The results of classification have shown a 35% improvement compared to classic MFCC feature in 50 dB noise level. In the work of (Beltrán-Márquez et al., 2012), spectral domain feature of cosine multi-band spectral entropy signature (CMBSES) is presented to succinctly represent the environmental audio signal. The CMBSES features of sound frames from nine classes are considered for recognition. The experiments are conducted in four different signal-to-noise ratios and the results show better performance can be achieved based on the proposed binary signature. However, the limitation of the approaches which based on purely spectral feature is that the temporal characteristic is not taken into account, this drawback could lead to decrease in recognition accuracy since environmental sounds are vary in time domain.

Considering the changing patterns of environmental sounds in time domain along with spectral information, (Karbasi et al., 2011) proposed a spectral dynamic feature to describe the dynamic information of spectral characteristic of environmental sound. Different from the first and second order derivatives of MFCC, the spectral dynamic features are based on the Fourier transform calculation of the Mel filterbank outputs from frames of a sound. The best classification rate in total of 90.06% is achieved by using the SVM classifier with linear kernel. Recently, (Sivasankaran and Prabhu, 2013) proposed an environmental sound classification algorithm by using spectral feature of sub-band energy and temporal feature of MP-based sparse decomposition coefficients. The applied features are combined with MFCC to generate the feature vector for classification. The experimental results show that the proposed features could provide a best accuracy of 95.6% in classifying sounds from 14 classes by using a gaussian mixture model (GMM). Generally, the above mentioned approaches have demonstrated that, it is important to combine spectral domain feature with temporal domain feature in order to gain a better accuracy in the recognition of environmental sound.

2.4. Audio-Visual Fusion for Heterogeneous Information of sound and image

Considering the information fusion research issue of artificial machines based on sound and image signals in recent years, though there have been a lot of works proved to be progressive in realizing autonomous object or event recognition for environment perception, most of them are still rely on the homogeneous features which collected from multi-sensors that provide the non-heterogeneous information for further fusion work, such as visual image and infrared image, data from sonar and laser radar, to name a few (see (Han et al., 2013b) and (Pérez Grassi et al., 2011) as examples).

It is obvious that, though image and sound signals are able to provide various information of surrounding environment alone, both of them have some limitations and shortcomings in perception compared to each other. For example, visual image is generated by the reflection of sun light, it is distinguishable because it consists of intuitive and unique representation of objects and the visual description is vivid and comprehensive because of the complexity in color, contrast and shape, except that visual image is very sensitive to obstacles, masks and lighting levels. At the same time, as a wave formed signal, sound is able to provide extra information of distance and location of the sound source as a highlight beyond the advantages of visual image information, and robust to obstacles compared with image. However, auditory data is not visualized information and need sophisticated computational model to process which lead to the limitation of distortion when noise exists. Consequently, the fusion of heterogeneous information of sound and image still has not been widely researched due to the lack of computational model and state of the art techniques. Currently, one popular application in audio-visual fusion research field is known as the subject of speaker identification and tracking in which auditory or acoustic information is used only as low level features or supplementary information in these works. For example, series of works in (Fisher et al., 2000) and (Fisher and Darrell, 2004) describe joint statistical models that represent the joint distribution of visual and auditory signals in which the maximally informative joint subspaces learning approach based on the concept of entropy for multi-media signal analysis is presented. This approach is purely a signal-level fusion technique, nonparametric statistical density modeling techniques are used to represent complex joint densities of projected signals and to characterize the mutual information between heterogeneous signals, thus indicate whether if the heterogeneous signals belong to a common source, or more specifically, a single user. The method is tested on simulated as well as the real data collected from real world, and experimental results show that significant audio signal enhancement and video source localization capability could be achieved by using the suggested approach. Another hybrid approach is presented in (Chu et al., 2004), in which a multi-stage fusion framework is applied to combine the advantages from both feature fusion and decision fusion methodologies by exploiting the complementarity between them, in which the information from auditory and visual domain are fused in decision level by using a multi-stream Hidden Markov Model (HMM). Indoor environment audio-visual speech database consists of full-face frontal video with uniform background and lighting is used for recognition experiment and the results show that the proposed fusion framework outperforms the conventional fusion system in human voice recognition processing. Nonetheless, the typical natural environmental sounds are not considered in this approach.

Considering the application scenarios in robotics, (Ruesch et al., 2008) introduce a multi-modal saliency based artificial attention system, in which the saliency principle is used to simulate the bottom-up pattern of human beings and act as the linking bridge between sound and image information. The system combines spatial visual and acoustic saliency maps on a single egocentric map to aggregate different sensory modalities on a continuous spherical surface and yield the final egocentric and heterogeneous saliency map, by using which, the gaze direction of the robot is adjusted toward the most salient location and exploration of the environment is carried out. Though the experimental results obtained from real world test show that the gaze and react of robot to salient audio-visual event shift automatically and naturally, the acoustic saliency information is used merely as the location of the sound event instead of the semantic knowledge that can indicate which of what is happening in the environment, and the research work of learning new objects or events is not yet presented. In (Li et al., 2012) a two stages audio-visual fusion approach for active speaker localization is presented, in which the fusion of audio-visual information is applied to avoid false lip movement in the first stage and a Gaussian fusion method is proposed to integrate the estimates from both modalities of audio and visual. This approach is tested in a human-machine interaction scenario in which a human-like head is deployed as the experimental platform, test results show that significant increased accuracy and robustness for speaker localization is achieved compared to the audio/video modality alone. However, the fusion approach is carried out merely in the decision level or rather, the knowledge level instead of feature level, while the recognition work of natural sound from surrounding environment is still not involved in this research.

In general, it should be emphasized that, although some research works which focus on the fusion of audio-visual information has been conducted in feature level or decision level (i.e. knowledge level), the fusion methodologies been developed are mostly based on the complementarity of heterogeneous information. In these works, the sound information is used as a local information or extra information that to be fused with the dominant image information in some application scenarios, while as global knowledge information in others, yet the substantial fusion works or practical approaches that regarding heterogeneous information as equally important sources for intelligent perception of environment are still not well researched. To the best of my knowledge, there has been no concrete work to be accomplished to realize artificial awareness and intelligent perception for machines in complex environment based on the fusion of heterogeneous information in an autonomous way.

2.5. Conclusion

The overview of this research field along with the related state of the art techniques that motivate my work are illustrated in general in this chapter. It has demonstrated the published important works and approaches that related to my research in threefold: A) the review of saliency-driven visual information processing techniques which applied for image signal perception; B) the review of auditory saliency detection models as well as state of the art approaches and the previous works that have done on the topic of environmental sound recognition, in which the recognition and classification of event based foreground sounds are the main research orientation; C) summarization of current audio-visual information fusion works in either feature level or decision (i.e. knowledge) level to present an illustration of global fusion methodology, applications related to robotics are specially emphasized. Some distinct approaches and observations are given literally to provide the general consideration of the motivation of my work, thus the discussion regarding the state of the art works are connected to the problems that are researched in this thesis.

Especially, three general observations need to be emphasized based on the above mentioned bibliography work. Firstly, saliency is the common principle which exists in selective attention mechanism of human beings for both visual and auditory perceptions. Compare with the perception process by using entire information of individually visual and auditory channels, aware of the most salient objects or events will lead to a more fast and simple procedure in the perception of the surrounding environment. In other words, saliency based visual and auditory signal processing could be the key point of realizing the heterogeneous information based artificial awareness ability for machines as the simulation of human perception system. However, the research issue of auditory saliency detection is more complex than image saliency detection because the composition of environmental sound is varies in both temporal and spectral domains, while current researches have shown that the saliency properties from both of these two domains should be taken into account for more accurate saliency detection.

Secondly, as sound and image information are completely heterogeneous, current researches indicate that the features for further processing of each channel should be extracted differently. The limitations of feature extraction from both channels exist in the same time but in different forms. Regarding the image feature extraction, though many features have been proposed and proved to be effective, the main limitation for image features is the high dimension problem of feature space. Therefore, approaches that could reduce the redundancy of feature space should be the prior consideration in the image part. On the other hand, despite various sound features including spectral, temporal and acoustic features have been presented or designed to achieve the recognition task, the recognition rate still could decrease because of the unstructured characteristic of environmental sound. Meanwhile, though the spectrogram based sound feature extraction and recognition approaches in some works like (Dennis et al., 2013b), (Souli and Lachiri, 2012b), (Lin et al., 2012) and (Khunarsal et al., 2013), to mention only a few, have shown that it is a more distinct way of analyzing the sound signal in visual modality, the recognition accuracy could be sharply decreased by using features from single feature domain or applying traditional acoustic features, especially when strong background noise exists. In previous mentioned environmental sound recognition works, higher recognition accuracy could be obtained by applying multiple features from different feature space, thus it is a possible solution to combine multi-scale features which represent the acoustic properties from multiple aspects of environmental sound to achieve better recognition performance.

Thirdly, though past research works which focus on the audio-visual fusion proved to be effective in literatures supported by the experimental results, most of them, like (Fisher III and Darrell, 2004) and (Ruesch et al., 2008) are processing sound and image as two heterogeneous data and the fusion approaches are mostly based on certain kind of computational model. As a result, the fusion process will involve many mathematics issues which cost huge computation resources and are not similar to the biologically processing procedure in human brain which is purely a clustering and learning process. Consequently, in order to explorer the bio-inspired approach for sound and image fusion, new approach needs to be researched.

Chapter 3. The Detection and Classification of Environmental Sound based on Auditory Saliency for Artificial Awareness

3.1. Introduction

To provide machines with artificial awareness ability for intelligent environment perception, the detection and recognition approaches of salient environmental sound are presented in this chapter in order to simulate the auditory awareness characteristic of humans. Since previous work on this topic has been rarely conducted, most of the related research works are either focus on the auditory saliency detection or on the environmental sound recognition. Therefore, the presented method can be seen as a combination of two individual works which are auditory saliency detection and salient environmental sound recognition, respectively, however they are highly connected.

Over the last decade, many approaches have been presented to successfully detect the auditory saliency property from sound signals by using a spectrogram based auditory saliency map. As mentioned in previous chapter, several auditory saliency models (e.g. (Kayser et al., 2005), (Kalinli and Narayanan, 2007) and (Duangudom and Anderson, 2007)) have been proposed to detect salient sound, in which the local and global contrast features from spectrographic image are considered with respect to the time-frequency property of sounds. Though these models proved to be effective, the image saliency features from spectrogram are not realistic representations due to the existence of background noise, not to mention that the detection accuracy could be interfered by overlapped salient sounds. To overcome the shortcomings of applying only image saliency features, improvement approaches (e.g. (Kaya and Elhilali, 2012), (Tsuchida and Cottrell, 2012) and (Kim et al., 2014)) have been proposed by using saliency features from multiple domains, including temporal and spectral. Although these mentioned methods proved to be effective, the sound data applied in their tests are simple sounds including single tones and speeches, and little environmental sound from real world which consists strong background noise has been considered.

Considering the research of environmental sound recognition, all the methods presented in previous works can be categorized into two parts according to their key contributions. The first part focus on the design of feature and the second part relates to the design of classification method. Compare to the design of classification method (see (Su et al., 2011) and (Zhang et al., 2011) as examples), many works have been done in designing various features of sound to achieve the environmental sound recognition task, such as time-frequency features (Guo et al., 2012), MFCC based features (Beritelli et al., 2008), spectrogram based features ((Dennis et al., 2013b) and (Souli et al., 2011)), MP features ((Chu et al., 2008), (Li et al., 2010) and (Yamakawa et al., 2011)) and the combination of particular features (Chu et al., 2009). It is general accepted by the mentioned works that MFCC based feature is the baseline feature which should be combined with other features to achieve higher recognition accuracy. However, though the image features derived from spectrogram and MP features derived from sparse representation provide to be effective and robust in classifying environmental sounds, both of these features could be influenced by the background noise and cost lots of computational resources.

Motivated by the shortcomings and limitations of previous research works from both auditory saliency detection and environmental sound recognition, a salient environmental sound detection and recognition approach is presented. It has two stages in which auditory saliency detection of environmental sound and salient sound recognition are included. The general objective of this approach is to accurately locate the salient environmental sound and accomplish the recognition task by applying proper features.

To be specific, in the auditory saliency detection stage, heterogeneous saliency features from acoustic channel and image channel are combined to form the auditory saliency map. The saliency feature from acoustic channel consists of saliency features from both spectral and temporal domain, which represented by the local maximum points from PSD and MFCC of environmental sound signal. To improve the detection accuracy of this approach and provide robustness in dealing with overlapped salient sounds, these saliency features are further verified by using a computational IOR model and a spectral saliency calculation algorithm. Meanwhile, in order to overcome the drawback of using traditional RGB color space, the image saliency feature is obtained by using the log-scale spectrogram in opponent color space. In addition, a heterogeneous saliency feature fusion algorithm is proposed to yield the final auditory saliency map for salient environmental sound detection.

Previous mentioned research works have shown that, environmental sounds are different from human speeches and music signals, which indicate that the recognition performance of environmental sound highly rely on the selection of audio features. Since environmental sounds are various in characteristics cross different categories but similar within the same category, a multi-scale features based classification approach is presented for environmental sound recognition. Its basic hypothesis is that, similar environmental sounds (e.g. from one category) share similar power energy distribution in certain frequency bands. Nevertheless, this power energy distribution is difficult to specifically describe in a mathematical way, because similar sounds could have slightly different distributions in local frequency bands while globally similar. Therefore, the fuzzy vector inspired from fuzzy set theory is considered to represent the fuzzy property of the power energy distribution of similar environmental sounds.

3.2. Overview of the Approach

The proposed approach in this chapter consists of two general processing units which are the auditory saliency detection unit and the recognition unit of salient environmental sound. The general structure of my approach is graphically illustrated in Fig. 4, in which the composition of each unit is detailed presented.

It can be seen from Figure 4 that, when the environmental sound is perceived and recorded, it will firstly be processed by the auditory saliency detection unit in order to extract the salient segment from the original sound data. Thereafter, the salient environmental sound segment is treated as input of the sound recognition unit, in which the features of salient environmental sound is calculated to accomplish the recognition task and the information of environment sound is obtained.

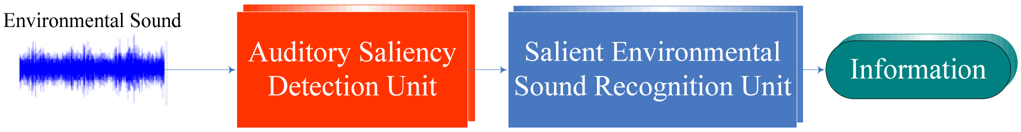

Environmental Sound
Auditory Saliency Detection Unit
Salient Environmental Sound Recognition Unit
Information

Figure 5: The overview structure of the presented approach.

In general, in the auditory saliency detection unit, multiple saliency features derived from both spectral and temporal domains are calculated and heterogeneous saliency features of acoustic and image channels are combined. In Figure 5, a block diagram of the auditory saliency detection unit shows the composition of particular modules in detail.

It is shown in Figure 5 that, the auditory saliency unit detects the auditory saliency property of input environmental sound in a parallel way. In module 1, the short-term Shannon entropy of the environmental sound signal will be calculated to provide global information for estimating the background noise level. According to the differential in background noise estimation, the time parameter of computational IOR model will be determined and applied in module 3 for temporal saliency feature verification. The salient sound segment obtained from module 1 contains the potential salient sound which indicated by the Shannon entropy, however, there do exists the possibility that the potential salient sounds which have been detected are background noises or multiple salient sounds. Therefore, the salient sound segment will be sent to module 3 and 4 for temporal and spectral saliency calculation to improve the accuracy of acoustic saliency detection. Note that, the local saliency features from temporal and spectral domains are calculated in model 3 and module 4, respectively, while image saliency feature from spectrogram is calculated in module 2 synchronously.

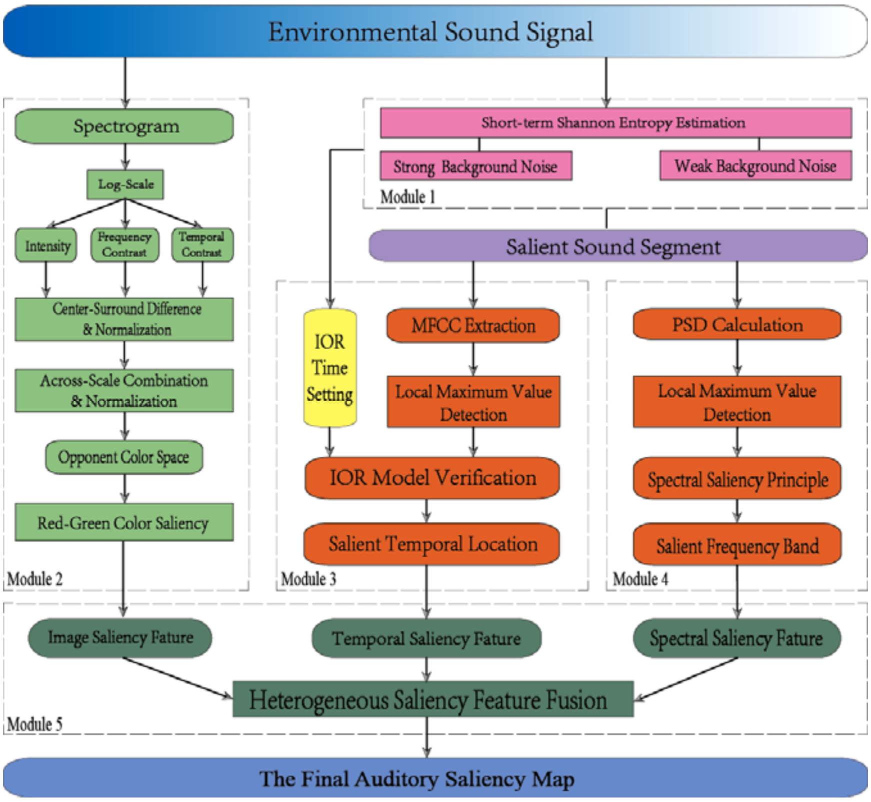


Figure 6: Block diagram of auditory saliency detection unit.

In module 3 and 4, both of the temporal and spectral saliency features are derived by detecting the local maximum value points in MFCC and PSD of the salient sound segment. To be specific, spectral saliency features are obtained in module 4 by using a designed spectral saliency principle to eliminate the pseudo salient points in PSD. Temporal saliency features are verified based on a computation IOR model which inspired by the inhibition of return phenomenon of human beings, in which the time parameter is different according to different background noise estimation. In module 2, the traditional spectrogram is applied to generate the auditory saliency map. However, since spectrogram could be easily contaminated by the background noise, a log-scale transformation is conducted to suppress the interference of lower level background noise and emphasize the salient sound signals. Meanwhile, as the classical RGB color space is not the best color representation for human perception, opponent color space is taken into account to improve the detection performance from spectrogram.

3.3. Heterogeneous Saliency Features Calculation

3.3.1. Background Noise Estimation

3.3.1.1. Shannon Entropy

Compare to the visual information perceived by mankind, sound will always exist and real silence is rare because the background noise is almost inevitable even in a tranquil environment. However, human beings are not bothered by this problem because only those sounds which are salient or relevant to one's expectations will be attended. In other words, the uncertain information caused by the salient auditory stimuli will trigger the perception as an instinct of self protection. In Shannon's information theory (Shannon, 2001), the concept of entropy was brought in to measure the uncertainty associated with a random variable. Since the salient sound could be seen as an uncertain signal source compared with its temporal neighborhood within a time period, the Shannon entropy could indicate the quantity of uncertainty information with any distribution contained by the salient sound.

3.3.1.2. Short-term Shannon Entropy

Since Shannon entropy is generally calculated by using the entire signal and could not reflect the variation tendency, I propose to calculate the short-term Shannon entropy (SSE) based on frames with 50% overlap to show the global change. Assume that s_j is the j^{th} frame of N samples of sound signal s, $s_{i,j}$ is the coefficients of j^{th} frame in i^{th} level, the SSE of one frame and entire signal are given as

$$E_j(s_j) = -\sum_i s_{i,j}^2 \log(s_{i,j}^2) \tag{1}$$

$$E_f(s) = \sum_{j=1}^{M} E_j(s_j) \tag{2}$$

The global saliency of sound signal s can be estimated as

$$G(s_t) = \begin{cases} \int E_f(s_t \omega(t-\tau)d\tau) & if \quad E_f(s_t) > \sigma_s \\ 0 & otherwise \end{cases} \tag{3}$$

In Eq. (3), σ_s is the threshold which determines the global saliency and $\omega(t)$ is window function for smoothing the discrete value. Hence, the large values of $G(s_t)$ represent the salient sounds. Meanwhile, the background noise estimation can be also given as

$$BGN_s = \begin{cases} 1, & if \quad 0_n G(s_t)/N < \theta_s \\ 0, & otherwise \end{cases} \tag{4}$$

Here, $0_nG(s_t)$ denotes the number of zero in $G(s_t)$ and N is the length of s_t, θ_s is the threshold of global saliency indicated by the SSE which experimentally set to be 0.4 in this thesis. The background noise estimator BGN_s equals 1 when the background noise is strong and indicates a weak background noise level when the value is 0. Note that, the noise status of strong and weak are referred to the global informative characteristic of sound instead of the true loudness. The unequal discriminant in Eq. (4) shows that, when θ_s×100% or less of the total number N are zero value points, the sound signal is non-zero value dominated which means that a great part of sound has uncertain information. Since informatics uncertainty represents the potential salient sound segment, the more non-zero values there are, the more potential salient segments exist.

The objective of Eq. (4) is to simply and efficiently estimate the background noise status of sound example for further processing. In general, sound examples recorded in most outdoor environment, especially in urban area, always contain eternal high level background noise, and the BGN_s will equals 1. Therefore, the saliency property of sound is based on the comparison with background noise. Especially, the short-term signal and sound with salient frequency components are the potential candidates for salient sounds. On the contrary, indoor or quiet outdoor environment is expected if the BGN_s equals 0, which means that the specific salient segment including overlapping sounds could be exist.

3.3.2. Temporal Saliency Feature Extraction

Inspired by human acoustic awareness characteristic, the auditory saliency can be seen as a temporal local contrast issue. As the computational representation of human hearing system, the coefficients of MFCC are used to represent the response of human beings to environmental sound signals. Therefore, the temporal saliency information could be obtained by finding the local maximum points of MFCC. However, since these points in MFCC are locally salient, not all of them indicate the real salient sounds. Thus, I propose a bio-inspired inhibition of return (IOR) computational model for saliency verification.

3.3.2.1. MFCC based Saliency Calculation

Related research works have shown that MFCC is very effective in simulating the acoustic characteristic of human hearing system. It is frequently used as the sound feature in many applications, such as human speech recognition, speaker identification and other audio signal related processing scenarios. MFCC is the calculation of cosine transform of short-term energy spectrum's real logarithm into mel-frequency scale, which is defined as follow

$$mel\left(f\right) = 2595 \cdot \lg\left(1 + f/700\right) \qquad \textbf{(5)}$$

The above Eq. (5) shows that the mel-scale frequency *mel*(*f*) corresponds to the real frequency in a nonlinear pattern. It has been widely accepted that the auditory characteristic of human beings could vary in different frequency bands, which means that we are more sensitive to low frequency signal than high frequency signal. In fact, the cochlea of our body can be considered as a filter bank that linearly response to the frequency lower than 1 kHz and nonlinearly response to the frequency higher than 1 kHz in a logarithmic manner. Thus, *mel*(*f*) is a better computational approximation of human hearing system in cochlea than linearly-spaced frequency bands because it simulates the nonlinear hearing sensitivity of human acoustic property to different sounds in different frequencies.

Meanwhile, *mel*(*f*) can also be seen as a subjective representation of the frequency feature of sound to the human awareness ability. In this thesis, *mel*(*f*) is implemented by a series of triangular band pass filters with scales of 12, and the MFCC is the result after converting the log mel spectrum into time domain. The calculation process of MFCC is showed in Fig. 6.

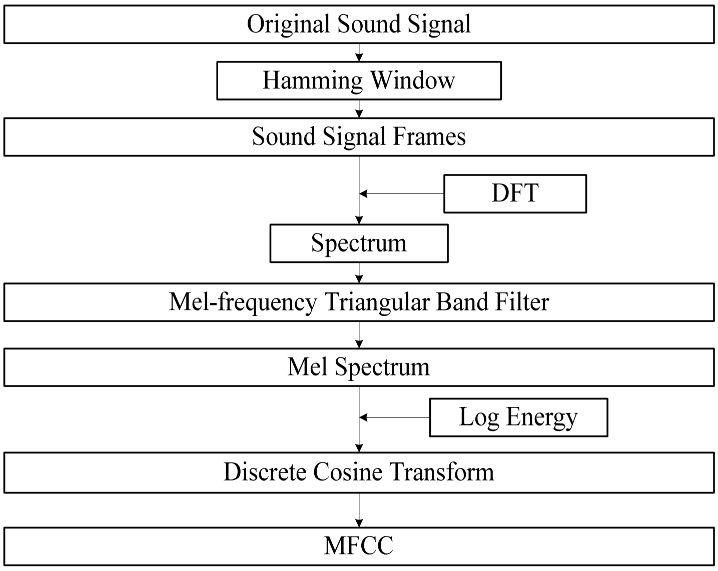


Figure 7: The calculation process of MFCC.

The temporal saliency feature of environmental sound is calculated by locating the local maximum value point in MFCC. Figure 7 demonstrates the detection result of a salient sound segment which derived from a real environmental sound example as an instance.

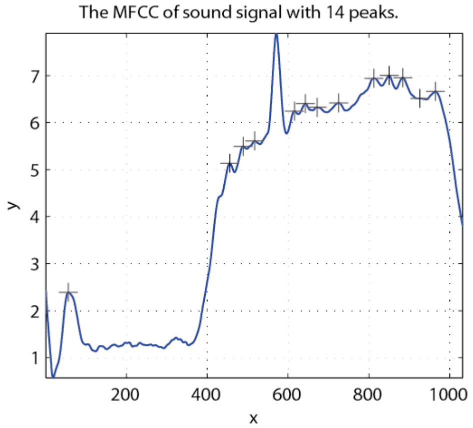


Figure 8: The temporal saliency indicated by the local maximum value point.

3.3.2.2. Computational IOR Model for Feature Verification

Human beings always intend to be attracted by the sounds with higher frequency components. Thus sounds with lower frequency components could be considered as pseudo background noise. Regarding the environmental sounds with low frequency components, the sound with salient loudness level among its neighborhood in temporal domain could also be perceived as salient sound. This saliency-driven selective attention mechanism can also be explained by the previous mentioned IOR phenomena which elucidates that new sound with a surprise value will be perceived more quickly than those sounds vary within a certain interval in spectral domain. This is the reason why we are sensitive to the spike-form short-term sound signal even a current salient sound has been attended.

To mimicking the acoustic saliency detection of human beings, the MFCC of sound is considered to be the main representation of human hearing system and the salient sound is indicated by the local maximum value of MFCC. However, there do exists the possibility that pseudo peaks in MFCC curve are irrelevant to the salient sound detection. An example is given in Fig. 7 to illustrate this drawback of purely mathematical approach. Therefore, I initially proposed a computational IOR model based saliency detection approach to locate the most salient sound in temporal domain.

Assuming there are *p* local maximum points of MFCC have been detected, the real salient points are the ones with a minimum time interval between their two neighbor points.

Psychology and neurobiology researches ((Abbott et al., 2012) and (Xu et al., 2014)) have shown that, stimuli with the stimulus onset asynchronies (SOA) within 200 ms are considered anticipatory response and facilitation effect occurred, while valid cues delay reaction times (RTs) relative to the invalid cues at longer SOAs (over 400 ms). This phenomena also supported by other research works (e.g. (Richards, 2000), (MacPherson et al., 2003) and (Gouzoulis-Mayfrank et al., 2006)) which experimentally demonstrated that, facilitation can be found with SOAs of 450 ms in infants or schizophrenics and between 360 and 570 ms in younger children. To simplified the calculation, the inhibitory effect of temporal saliency can be determined by using a minimum time interval t_{IOR} defined as

$$t_{IOR} = \begin{cases} 200ms, & if \quad BGN_s = 1 \\ 450ms, & if \quad BGN_s = 0 \end{cases} \tag{6}$$

In Eq. (6), the general assumption is that two local salient points within t_{IOR} will be considered as non-salient points. This can be explained by the IOR mechanism that, if the time interval between two temporal adjoining local salient points is less than t_{IOR}, facilitation will occurred and inhibitory effect is not obvious, so the latter salient point will be treated as non-surprise point by auditory perception system. Hence, the local salient points will be processed as current background noise. However, it is also possible that the local salient points represent a real salient sound globally, which means that they could be the reflection of local change in frequency patterns. Therefore, by combining other saliency features, the local salient points can be considered as pseudo background noise and a more salient sound is expected.

Since background noise is a crucial factor to salient sound detection, it is necessary to model IOR mechanism according to the differential in background noise level. Biologically speaking, when we exposed to multiple sounds as well as background noise, both sounds and noises will be perceived as input stimuli at the same time. If the background noise level is strong, the caused facilitation effect will be in a higher level, and sounds within t_{IOR} will be masked by the noise. Conversely, if the background noise level is weak, the auditory perception system will be more sensitive to non-noise sounds. Meanwhile, due to a lower level of facilitation effect, the auditory perception will be influenced by overlapped salient sounds, in which facilitation will caused by the less salient sounds. Thus, the quasi-salient sounds could be treated as pseudo background noise and the more salient sounds are expected.

Moreover, as strong background noise can be seen as multiple temporal stimuli that closed to each other, here I choose the general accepted minimum time length of SOAs of 200 ms to represent the facilitation process. Consequently, those separate stimuli in sound longer than 200 ms are the candidates of real salient sounds. Regarding the sounds under weak background noise, since facilitation caused by noise will be in a lower level, we could be more affected by the facilitation generated by the quasi-salient sounds. In other words, our auditory perception system is intend to perceive new sounds even a current salient sound has been attended. From a biology point of view, the most salient sound in overlapped period to human auditory perception is mostly the short term separate sound signal, which means that the long term salient sounds would be frequently considered as the pseudo background noise. According to the neurobiology researches in IOR phenomena, I experimentally set the t_{IOR} to be 450 ms in weak background noise environment.

In addition to the mentioned temporal requirement of IOR model, it is assumed that the detected point should also have a salient value of MFCC even if it has satisfied Eq. (6) already. The mechanism of IOR shows that, under the circumstance that overlapped salient sounds exist, if one salient sound occurred and attended, the auditory perception system is expecting or more sensitive to a more salient sound. On the other hand, the decrease of salient MFCC value also indicates the saliency properties of sounds. To be specific, for two adjoining salient points which has meet the requirement of Eq. (6) if the first salient point has been assigned as candidate of the most salient sound, there should be a significant decrease between the MFCC value of these two points. Otherwise, the candidate does not represent the real salient sound and probably is the local saliency representation of pseudo background noise. Therefore, the expectation value of MFCC for *p* point can be calculated as

$$M_{IOR+,p+1} = M_p \left(e^{t/t_{IOR}} \alpha_s \right)^{-1} \tag{7}$$

$$M_{IOR-,p+1} = M_p \left(e^{t/t_{IOR}} \beta_s \right) \tag{8}$$

where M_p is the MFCC value of *p* point, *t* is the time length between *p* salient point and *p+1* salient point, α_s and β_s are the adjustment coefficients. It is obviously that $M_{IOR+,p+1} > M_{IOR-,p+1}$. $M_{IOR+,p+1}$ and $M_{IOR-,p+1}$ are the two threshold values, which respectively represent the upper limit and the lower limit of the expectation value for the detected *p+1* salient point. Eq. (7) and Eq. (8) simulate the saliency detection principle that influenced by the IOR mechanism, in which the initiate value of M_{IOR+} and M_{IOR-} are high and will decrease as time lapses. Furthermore, considering the existence of overlapping, the most salient sound of *p* point can be detected by the inequable discriminator if

$$M_p \geq M_{IOR+,p+1} \cap M_{p+1} \leq M_{IOR-,p} \tag{9}$$

Consequently, the most salient sound can be detected among background noise or other relatively salient sounds.

3.3.3. Spectral Saliency Feature Extraction

3.3.3.1. Power Spectral Density

Several auditory saliency models have been proposed to reveal the saliency characteristic of sound and proved to be effective as mentioned in previous section. However, most saliency features are derived from local contrast while global saliency information is rarely taken into account. Therefore, I propose to take global saliency feature from the power spectral density (PSD) estimation of a given sound signal to obtain the salient frequency component distribution as a complementary part of the traditional auditory saliency model.

The PSD calculation process of a given sound signal is that, divide the sound signal into several overlapping segments by applying a window function to each segment and then averaging the periodogram of each segment thereafter. As the estimation result indicates the power distribution of signal at different frequencies, the local maximum point in the spectrum curve can be used to locate the salient components of sound signal in spectral domain.

3.3.3.2. PSD based Saliency Calculation

Assuming that the local maximum value points has been detected in the curve of PSD which indicate the potential salient frequency components, I compare each point with P_{mean} which is the mean value of PSD and ignore the pseudo salient point that less than P_{mean}. Meanwhile, the frequency range of human hearing system is conventionally on average from 20 Hz to 20 kHz, but the lower and upper limits of this range can only be heard to very few people. To most of the adult individuals, the audible sounds in real world are with frequencies from 40 Hz to 16 kHz in an unequal sensitivity pattern and great sensitivity can be achieved normally in the frequency range 2 kHz to 5 kHz. Furthermore, two sounds with different loudness can be distinguished if the physical level increases by 10 dB (Petit et al., 2013). Therefore, the salient frequency components that above 15 kHz will be reassigned as non salient, and for the rest q salient points, the frequency axis of PSD will be divided into $q+1$ bands as $(0, f_1), (f_1, f_2), \ldots, (f_{q-1}, f_q), (f_q, f_{Pmean})$, where f_q represents the location of the q salient point in frequency axis and f_{Pmean} is the location of frequency's mean value point. Then the local minimum value $P_{local\min,j}$ in the j^{th} band can be found and check the saliency by using the following inequal discrimination as

$$S_{PSD,q} = \begin{cases} 1, & if \quad P_q - \forall\left(P_{local\min,j}, P_{local\min,j+1}\right) > \zeta \\ 0, & otherwise \end{cases} \tag{10}$$

Here, ζ is the saliency distance that equals 10 dB, P_q is the PSD value of the q salient point and $S_{PSD,q}$ indicates the true saliency property of q point. Therefore, those salient points in PSD with $S_{PSD,q}$=1 are the real salient frequency components with more power spectral density energy. They will be processed in the conventional auditory saliency map within a spectral period of (f_q ± 500Hz) to decrease the computational cost. Nevertheless, if $S_{PSD,q}$=0 for all the q points, it means that no salient frequency component exists. Thus, there is a great possibility that the background noise is too strong or the sound signal has widely distributed frequency component, then the temporal saliency features play a vital role in the saliency detection process.

3.3.4. Image Saliency Detection from Spectrogram

The image saliency feature is derived by using the opponent color space based on the log-scale spectrogram, in which the color contrast of red-green channel represents the time-frequency saliency property that visually demonstrated by the spectrogram. Meanwhile, the purpose of using log-scale spectrogram is to depress the influence of background noise to the original spectrogram and emphasize the potential salient signal.

3.3.4.1. Log-Scale Spectrogram

Firstly, to depress the affect of background noise and emphasize the salient sound signal components, the original spectrogram is transformed into log scale. Assuming that the original spectrogram is I_s, the log scale spectrogram I_g is calculated as

$$I_g = 20\log_{10}(|I_s|^2 / 60) \qquad \textbf{(11)}$$

In I_g based time-frequency representation of sound, signal components with higher power energy will be represented by the red region in RGB color space while lower power components are commonly in green.

3.3.4.2. Image Saliency Calculation based on Opponent Color Space

After the logarithmic transform, the value of each point in spectrogram *I* has been changed while the color is still represented in RGB color space. However, as salient signals that with strong power are always in red color and lower power signals are generally in green, the opponent color space is considered to have better representation ability (Evangelopoulos et al., 2008). Therefore, the image saliency feature from I_g can be calculated as $I_{r\text{-}g} = I_{g(red)}\text{-}I_{g(green)}$, in which the $I_{r\text{-}g}$ is the saliency information derived from opponent red-green channel, in which $I_{g(red)}$ and $I_{g(green)}$ represent the red and green channel information from I_g respectively.

However, since the human perception of color is not best represented in RGB color space according to the opponent-process theory, a better model of opponent color space is proposed in which the red and green colors are postulated as opponent colors. Hence, as the salient time-frequency components of sound signals with red color are more salient to human vision among background or spatiotemporal neighborhood with green color, the image saliency can be derived from the red-green channel (Anwer et al., 2011) of opponent color space which defined as

$$I_{red-green} = \left(I_g \left(RGB_r \right) - I_g \left(RGB_g \right) \right) \tag{12}$$

where the RGB_r and RGB_g are the red and green channel of original RGB color space, respectively. As a result, a spatiotemporal saliency map based on spectrogram is obtained by combining the auditory saliency feature of salient Mt from temporal domain and saliency feature from spectral domain of $S_{PSD,q}$ together.

3.3.5. Heterogeneous Saliency Feature Fusion

By combining the above calculated heterogeneous saliency features together, the final auditory saliency is achieved by a fusion process which given as

$$S_{image} = I_{r-g} \cap \left\{ S_{SSE} \right\} \qquad \textbf{(13)}$$

$$S_{spectral-temporal} = \begin{cases} \left\{ S_{M_p} \right\} \cup \left\{ S_{PSD,q} \right\}, & if \quad S_{PSD,q} \neq \varnothing \\ \left\{ S_{M_p} \right\} \cap \left\{ S_{P>P_{mean}} \right\}, & if \quad S_{PSD,q} = \varnothing \end{cases} \qquad \textbf{(14)}$$

$$S_{af} = \begin{cases} S_{image} \oplus S_{spectral-temporal}, & if \quad BGN_s = 1 \\ S_{image} \oplus (S_{image} \,||\, S_{spectral-temporal}), & if \quad BGN_s = 0 \end{cases} \qquad \textbf{(15)}$$

Here, the set $\{S_{SSE}\}$ in Eq. (13) represents the salient segments of the sound which indicated by $G(s_t)\neq 0$, the set $\{S_{Mp}\}$ in Eq. (14) contains the temporal locations of salient signals, and the set $\{S_{P>Pmean}\}$ represents the salient frequency components of which PSD values are greater than P_{mean}. In Eq. (15), the final auditory saliency is obtained by fusing image saliency feature and the spectral-temporal saliency feature. The symbol of '$\oplus$' in Eq. (15) means that two saliency features should be fused with each other in an equal important way that similar to the logic and operation. The symbol of '||' means that the fusion principle is more close to the logical or operation that spectral-temporal saliency feature will be used as a constraint in a dominant way. Therefore, the spectral-temporal saliency feature should be fused with image saliency feature first as a constraint when BGN_s=0.

3.4. Multi-Scale Feature based Salient Environmental Sound Recognition

3.4.1. General Introduction

Regarding the environmental sounds that happened in most of the application scenarios, human acoustic awareness ability is able to distinguish them by using the spectral energy feature and temporal characteristic. Two sounds with similar spectral energy distributions could be recognized into one general class and temporal features are combined afterwards to yield a more accurate recognition result. To simulate the fuzzy recognition ability of human awareness, a two-level sound classification approach is presented, in which multi-scale features from spectral and temporal domain are applied.

3.4.2. Multi-Scale Feature Selection

So far, many research works have shown that the MFCC coefficients based acoustic features are effective and efficient in speech or music signal recognition. Regarding the topic of environmental sound recognition, other features are frequently considered as an supplementary part of the traditional MFCC features to achieve higher recognition accuracy. Nevertheless, despite that complex features can provide higher recognition accuracy, it is obvious that the computational cost will increase. Moreover, for the new type of sounds that similar to each other within the same category, recognition accuracy could be sharply decrease and recogntion could even be failed if the input sound has no pretrained models or templates to classify.

However, human beings are able to recognize different sounds from different categories by using the spectral characteristics. For example, when the sounds of alarm and dog barking are percieved by auditory perception system, they could be easily distinguished according to the different spectral properties, such as the distribution of specific frequency components and particular frequency bands. Meanwhile, since different typies of sounds from the same category could be slightly different in spectral properties according to different scenarios, spectral energy distribution of the same sound could also be different locally while be similar globally. Therefore, it is necessary to develop a feature that can fuzzily describe the general similarity in the spectral energy distribution of sounds which from the same category.

Therefore, by applying the fuzzy representation from fuzzy set theory (FST), a novel audio feature of fuzzy vector which based on the spectral energy distribution is introduced. This spectral feature is combined with MFCC based acoustic features to yield the multi-scale features for environmental sound recognition work. To be specific, in the proposed two-level classification process, the first level is based on the fuzzy vector of spectral energy distribution of sound as coarse classification. The second level is the finer classification in which combination of MFCC, delta-MFCC and delta-delta-MFCC are considered as the temporal features.

3.4.2.1. Fuzzy Set Theory

Since the initially formalization by (Zadeh, 1965) in 1965 as an extension of the classical notion of set, the fuzzy set theory has been widely used in many applications (see (Zimmermann, 2010) as a review). The notion of fuzzy set theory is proposed to accommodate fuzziness in the sense that it is contained in human language, that is, in human understanding, evaluation, and decisions. From mathematical point of view, the fuzzy set theory is proposed to represent uncertainty and vagueness of data, and to provide useful tools for processing imprecision information. Conventional approaches which based on the classical sets lack the means for representating knowledge of fuzzy concepts, however, fuzzy set theory is able to provide strict mathematical methodology in which vague conceptual phenomena or desciption can be precisely and rigorously studied.

For example, in describing the temperature of envrionment or weather, we often use the words like ‘cold’, ‘warm’ and ‘hot’ instead of particular degrees. However, it is difficult to clearly define the relationship between words and specific temperature by using classical sets, because opinion could be completely different as to whether 25 degrees Celsius is ‘warm’ or ‘hot’ according to different people. Meanwhile, it is also a difficult task to define the transiton from ‘cool’ to ‘warm’ by applying the measurement of degree Celsius, which makes it a mathematically singularity to determine the membership function between lexically knowledge and specifica value. However, this limitation could be overcomed by using fuzzy set theory and more accurate description of natural phenomenon or fuzzy concept could be achieved. In Figure 8 an example has been given to show how fuzzy set descibes this natural fuzzy knowledge.

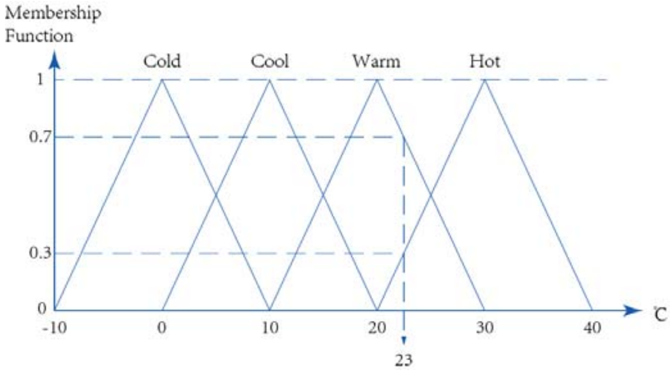


Figure 9: The membership function for temperature (Adapted from (**Majozi and Zhu, 2005**)).

According to the definition of fuzzy set, a fuzzy set A can be defined as

$$A=\left\{\left(x_f,\mu_A\left(x_f\right)\right)|\,x_f\in X_f\right\} \quad \textbf{(16)}$$

where X_f denotes the collection of all the objects and be a nonempty set, x_f represents the generally element of the fuzzy set, $\mu_A(x_f)$ is the membership function of x_f which can be characterized as

$$\mu_A : X_f \rightarrow [0,1] \quad \textbf{(17)}$$

Considering the general application in real world, the description of one object or event could not be infinite length. Therefore, if $X_f = \{x_{f1}, \dots, x_{fn}\}$ is a finite set, a more frequently applied form of fuzzy set can be notated as

$$A=\mu_A\left(x_{f1}\right)/\,x_{f1}+\cdots+\mu_A\left(x_{fn}\right)/\,x_{fn}=\sum_{i=1}^{n}\mu_A\left(x_{fi}\right)/\,x_{fi} \quad \textbf{(18)}$$

where $\mu_A\left(x_{fi}\right)/\,x_{fi}, i=1,\dots,n$ is not a subtraction operation but signifies that the $\mu_A\left(x_{fi}\right)$ is the grade of membership of x_{fi} in A. The plus " + " sign is also not the actual plus operator but represents the unification of all the membership grades.

3.4.2.2. Fuzzy Vector based Feature Extraction

While environmental sounds have different spectral energy distributions, some particular frequency bands could be shared by several kinds of sound. Since the hearable sounds to human auditory perception are normally from 20 Hz to 20 kHz, thus the entire hearing range could be considered as object domain. To simple the calculation, I propose to divide the frequency bands that sensitive to human auditory awareness which less than 20kHz into four subband as *f1 = (0, 5kHz)*, *f2 = (5kHz,10kHz)*,*f3 = (10kHz, 15kHz)* and*f4 = (15kHz, 20kHz)*.

As the spectral energy of most sounds will concentrate on particular subbands, the principal spectral component of sound can be calculated as $P_s=\lambda_p P_{mean}$ where the P_{mean} can be obtained from previous section, λ_p is a adjustment factor that determines the energy ratio of principal spectral component and emprically set to be 0.6 in this thesis. Then the spectral energy distribution of a sound signal can be represented by the proportion of energy that higher then P_s in each subband in the form of fuzzy vector (FV) as

$$A_s = \mu_1 f_1 + \mu_2 f_2 + \mu_3 f_3 + \mu_4 f_4 \quad \textbf{(19)}$$

where μ_i are the energy ratio in each frequency band that higher than P_s.

Consequently, the FV of $(\mu_1, \mu_2, \mu_3, \mu_4)$ could be obtained from Eq. (19) for each sound signal, which can be considered as the spectral features of environmental sounds for classification

3.4.2.3. Acoustic Features Calculation

In addition to the classical MFCC feature which purpose is to represent the static characteristics of environmental sounds, the first and second derivatives (Δ's and ΔΔ's) are computed to represent the dynamic characteristics of sounds. The calculation of these two features can be given as

$$d_i = \sum_{n=1}^{N} n\left(c_{n+i} - c_{n-i}\right) / 2\sum_{n=1}^{N} n^2 \tag{20}$$

where d_i is the delta coefficient of frame i, c_{n+i} and c_{n-i} are the basic coefficients. The delta-delta coefficient can be obtained by replacing the basic coefficient with d_i.

3.4.3. Classification Approach

For the environmental sound classification, I propose to use SVM because all the derived features are in the form of vectors, and the numbers of class as well as data per class are small. In this thesis, the SVM classifier is trained by using the toolbox of LIBSVM which will be discussed in section 4.4.3 in detail. Particularly, in both of the two classification processes, two SVM classifiers are trained by using fuzzy vector based spectral features and MFCC based temporal features of sound example, respectively. Especially, the MFCC based features include 13 basic coefficients, 13 delta coefficients as well as 13 delta-delta coefficients. The training process use training data to classify environmental sounds into their respective classes, while the training data need to be sufficient to be statistically significant. Then, the SVM learning algorithm is applied to produce the classification parameters according to calculated features which used to classify test environmental sounds data afterwards. After the first level classification, the MFCC based features will be calculated for the examples that are not correctly classified, thus the second level classification is conducted to yield the final classification results.

3.5. Experiments

3.5.1. Validation of Salient Environmental Sound Detection

3.5.1.1. Data Setup

In order to verify the performance of proposed auditory saliency detection approach in dealing with sound signal occurred in real environment, two sound examples recorded in real outdoor environment are used. Each example contains at least one kind of sound occurred in everyday life which are salient to human awareness. Meanwhile, the properties of salient sounds contained in the sound examples are vary in both spectral and temporal domain. To be specific, example A recorded the sound event of police car deploying in which the salient sounds contain a siren of police car with two different frequency patterns and a sudden beep. Example B is a record of festival march in which the salient sounds to human hearing awareness are the multiple sounds of the horse's hoof hitting the ground. The background noises are included in both of the sound examples but vary in the loudness level. Respectively, the background noise exists in example A is rarely salient to the police siren, while the difficulty of auditory saliency is to distinguish the sudden beep of truck from the salient sound of police siren as background noise. The background noise exists in example B is at a very high level and almost as strong as the sounds of horse's hoof hitting the ground which are salient to human awareness.

The most salient sound of truck beep from example A and the salient sound from example B are typical short-term sound signals which could be masked by the background noise or any less salient sounds. The frame length of SSE is 1024 points with an overlap of 512 points and the scales of mel-scale filter bank are 20. The coefficients α_s and β_s in Eq. (7) and Eq. (8) are empirically set to 0.133 and 0.114, respectively.

3.5.1.2. Experimental Protocol

Based on the proposed auditory saliency detection approach, the two examples will be processed by using the MATLAB software on a laptop to obtain the saliency map, thus the performance of salient environment sound extraction can be validated.

Particularly, the protocol of experiment can be illustrated as the following processing procedure:

1) For each sound example s, calculate the short-term Shannon entropy (SSE).

First, divide *s* into small frames by a window length of 1024 point with overlap of 512 point. For each small frame, calculate the wavelet packet Shannon entropy and merge them together to form the SSE for entire signal. Normalize SSE to the interval of (0,1).

2) Estimate the BGN_s of sound example *s* and obtain the corresponding t_{IOR}.

For the SSE of entire signal, perform smoothing operation by applying a moving average of 30 point span. Set the segments with value smaller than σ_s to be zero so as to obtain the salient segments. Calculate the ratio of zero value point and compare it with θ_s×100%, output the BGN_s for sound example *s*. Set the corresponding t_{IOR} for sound example *s*.

3) Calculate the temporal saliency from MFCC of sound example *s*.

Calculate MFCC of sound example *s*, perform a smoothing operation by using the local regression method of 30 point span in which weighted linear least squares and a 1st degree polynomial model are applied. Find the local maximum value points M_{mx} of MFCC, calculate the time length between two M_{mx} points.

4) Perform the temporal saliency check by using the computational IOR model.

Start from the first M_{mx}, compare the time length between itself and latter point. If the time length is smaller than t_{IOR}, the latter M_{mx} point will be eliminated. For the remained M_{mx} points, calculate the upper and lower limits of $M_{IOR+,p+1}$ and $M_{IOR-,p+1}$ of each point. For each remained M_{mx} point, its MFCC value should larger than the upper limit $M_{IOR+,p+1}$ derived from the previous M_{mx} point (true neighborhood), or the MFCC value of latter point should be smaller than the lower limit $M_{IOR-,p+1}$ of current M_{mx} point, otherwise it should be eliminated. The M_{mx} point that pass the validation of computational IOR model is (or are) the real salient point (points).

5) Calculate the spectral saliency from PSD of sound example *s*.

Perform the Welch method for power spectral density estimation. Set the window function to be a Hamming window with 64 points, number of FFT to be 1024 points with the overlap of 50 points. The frequency range is set to be ‘oneside’ as only the real-valued signal is considered in this thesis. Find the local maximum value points P_{mx} of PSD and the local minimum value point $P_{local\min}$ of each frequency band that divided by all the P_{mx} points, calculate the difference values between each P_{mx} point and its two neighbor $P_{local\min}$ points. Compare the two difference values of each P_{mx} with 10, eliminate the P_{mx} points which have no difference value that larger than 10.

6) Generate the visual representation of sound *s* in log scale.

For sound example *s*, generate the spectrogram which can be considered as the visual representation of sound for saliency detection. To calculate the spectrogram I_s, the length of Hamming window is set to be 512 points with an overlap of 256 points, the number of FFT is set to be 512. The estimation of short-term, time-localized frequency content of sound *s* is applied in the log scale transformation to obtain the log scale short-time Fourier transform result. The obtained result is then meshed with frequency and time to generate the log scale spectrogram I_g.

7) Calculate image saliency in opponent color space.

The visual saliency is calculated by using the information from red-green channel from opponent color space of I_g. To be specific, the saliency property is calculated by making the value from red channel subtract the value from green channel.

8) Fuse the heterogeneous saliency feature to obtain the final saliency map.

Firstly, for the salient segments indicated by the SSE which $G(s_t)\neq 0$, the temporal locations of which are correlated with columns in I_g, which means that the column pixels correspond to the temporal positions of salient segments should be keep while other column pixels should be eliminated. Second, if there is no salient point of PSD, the corresponding column pixels in I_g that correlate with the temporal position of M_{mx} should be merged with the row pixels that correlate with the frequency band which PSD value is larger than the mean value in an union operation. However, if there are salient points of PSD, the corresponding column pixels in I_g that correlate with the temporal position of M_{mx} should be merged with the row pixels that correlate with the salient frequency bands (f_q±500Hz) in an intersection operation. Third, if the BGN_s=1, the saliency maps obtained in the first and second steps should be fused in a logic plus way. If the BGNs=0, the saliency maps obtained in the first and second steps should be fused first in a logic and way, then fused with the saliency map obtained in the first step in a logic plus way.

3.5.1.3. Verification Results and Discussion

The result images from multiple stages of auditory saliency detection process are graphically illustrated in Figure 9 to Figure 12, in which the results correlate to example A and B are organized in column (a) and (b) of each image, respectively. To compare the performance of proposed saliency detection approach, final auditory detection results derived from presented approach and the conventional approach of Kayser's work are given in Figure 13.

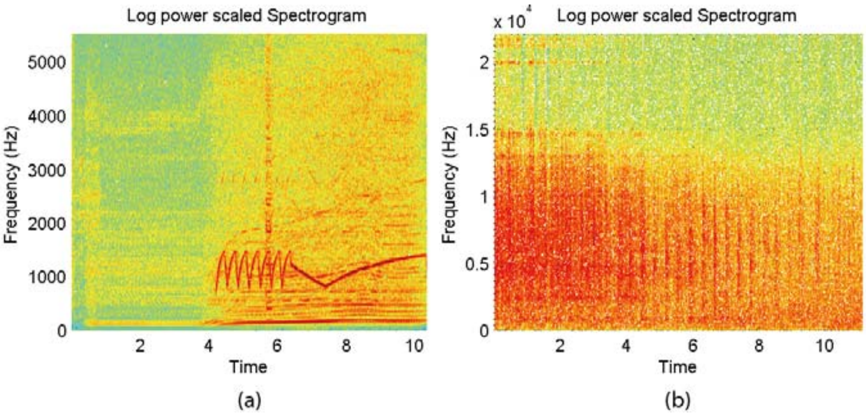


Figure 10: The log-scale spectrogram of environmental sounds.

The original spectrograms in log scale of the two sound examples are presented in Figure 9, in which the salient sound signals are represented by the red color regions. In the left column (e.g. (a)) of Figure 9, three salient sounds to human auditory awareness are distinguishable for human visual awareness in the log-scale spectrogram, in which the polygonal lines with salient red color in the low frequency region of spectrogram indicate the two salient sounds with different spectral patterns and a salient vertical line with red color which is salient to its neighborhood indicates the sudden beep of the truck in sound example A. The image from the right column of Figure 9 shows the time-frequency representation of sound example B. It is clearly to see that, despite that the spectrogram has already been transformed into log-scale, almost the entire spectrogram which less than 15 kHz in frequency axis is still in red color because of the strong background noise. Therefore, the salient sounds of horse's hoof hitting the ground are not possible to be accurately detected for human visual awareness due to the interference of background noise. However, the multiple salient sounds could be distinguished by human beings with auditory awareness.

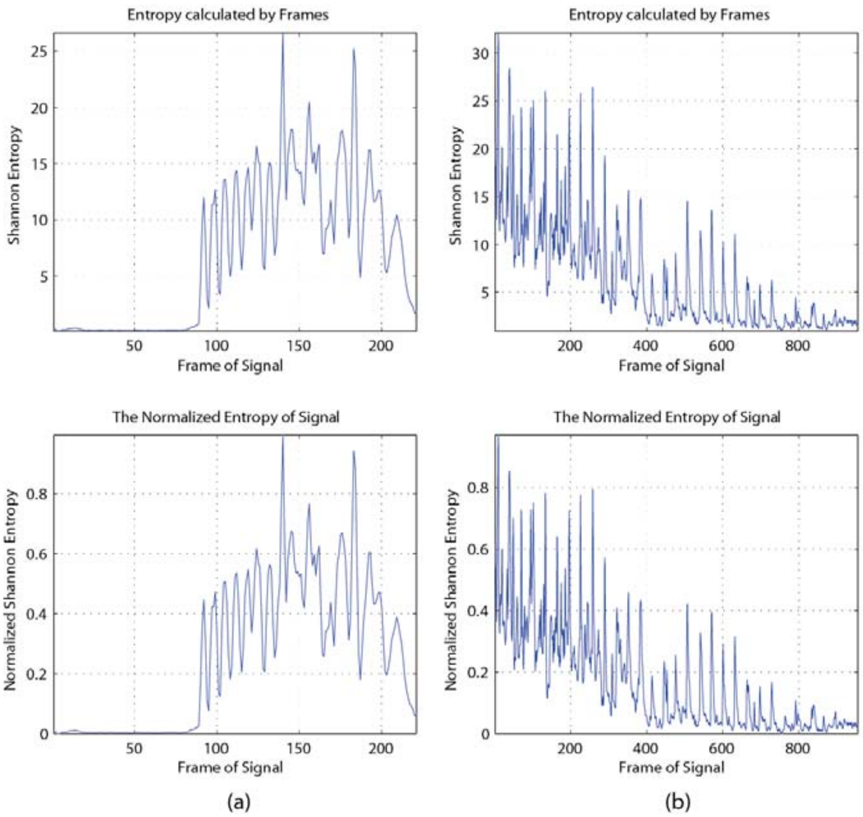


Figure 11: The short-term Shannon entropy and normalization of environmental sounds.

Images from Figure 10 show the SSE as well as the normalized SSE calculation results of sound example A and B are demonstrated in the column (a) and column (b), respectively. The left column images of Figure 10(a) show that there is a long period of which the values of SSE and normalized SSE are close to zero. This indicates that the degree of informatics uncertainty in example A is not always high, which infers that the global background noise should be in a lower level and segments of SSE with values over threshold θ_s could represent the potential salient sounds. Consequently, only the segments with values over θ_s of SSE and normalized SSE which represent the global temporal locations of salient sounds are allowed for further processing. Therefore, the computational cost could be decreased without reducing the detection performance. The SSE and normalized SSE of example B are presented in column (b) of Figure 10. The result images show that almost the entire sound example has the saliency property because the informatics uncertainty degrees of majority segments of the sound example are in a higher level (e.g. over threshold θ_s). Thus, further saliency detection from other aspects is required to achieve better accuracy.

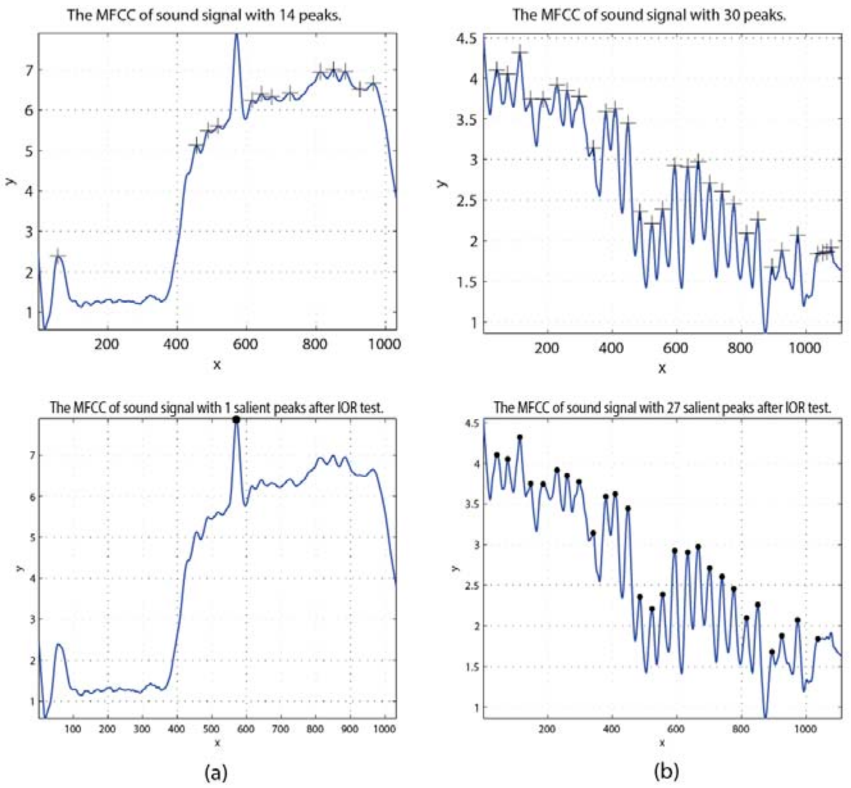


Figure 12: The local maximum value points of MFCC and the saliency detection results after IOR test.

Figure 11 illustrates the preliminary temporal saliency detection results based on the MFCC of the two experiment sound examples, in which the plus sign indicates the location of local maximum value point. It is obvious to see from the first row of Figure 11 that, multiple local maximum value points have been detected. However, to human hearing awareness, only one plus sign (the fifth from left) in the first image of column (a) from Figure 11 represents the temporal location of most salient sound while others are mismatched points. The first image of column (b) from Figure 11 also shows that most of the salient sound signals have been located in MFCC, while the last three salient points are yet mismatched according to human auditory saliency property.

This is because that all the salient points are locally maximum points and could not provide global saliency information of sound. To improve the detection accuracy, the computational IOR model presented in section 3.3.2.2 is applied and the results are given in the second row of Figure 11. The black dot in the image of second row from Figure 11(a) shows that, the most salient sound has been accurately detected after the IOR test, and other mismatched points which represent the pseudo salient background have been ignored. Relatively, the temporal saliency detection result of example B is presented in the image of second row from Figure 11(b), in which the mismatched salient points at the end of the sound example have also been removed.

The images from Figure 12 demonstrate the spectral saliency detection process, in which images from the first row show the preliminary results and final detection results are presented in the images of second row. Both the images from the first row of column (a) and (b) in Figure 12 show that, several local maximum value points are located and the value of P_{mean} is labeled. However, compare to the human auditory awareness characteristic, most of the local maximum value points detected in PSD curves of both sound examples are mismatched. Specifically, in the image of the first row from Figure 12(a), only one local maximum value point (the third from left) represents the real salient spectral component. While in the image of the first row from Figure 12(b), there is no real salient component exists because the values of PSD which are more than P_{mean} have no global maximum point. As a result, the preliminary saliency detection result in the image of first row from Figure 12(b) has no corresponding relationship with the spectral saliency feature of salient sound, which brings in the requirement of further saliency verification process if the salient sounds have no specific salient frequency band.

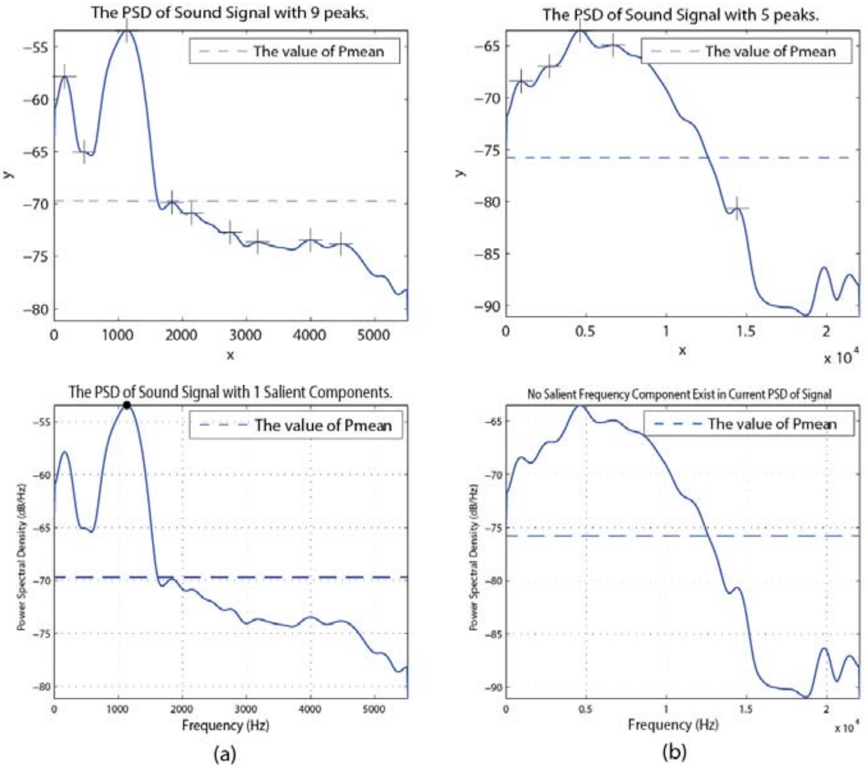


Figure 13: The spectral saliency detection results of sound examples.

The authentic spectral saliency feature detection results are demonstrated in the images of second row from Figure 12, in which the second image of Figure 12(a) shows that the salient frequency component has been correctly detected and located. Meanwhile, the local maximum value point which correlates to the background noise at lower frequency band is ignored. Regarding the result of example B which showed in the image of second row from Figure 12(b), though there are several individually local salient points have been detected previously, no authentic salient point exists after using the spectral saliency verification equation of Eq. (10). Hence, it indicates that frequency components with PSD value that over P_{mean} should be entirely considered as the spectral saliency feature for detection.

The final auditory saliency detection results are given in Figure 13 in the form of auditory saliency map. The corresponding detection results derived from Kayser's work are presented in the second row of Figure 13 for comparison. It is shown that, the salient sounds in both of the two examples have been successfully detected and clearly represented by the auditory saliency maps. To be specific, as shown in the image of the first row from Figure 13(a), the overlapped salient sound of the sudden beep of truck in example A has been accurately detected among the pseudo salient background which consists of two patterns of police car siren.

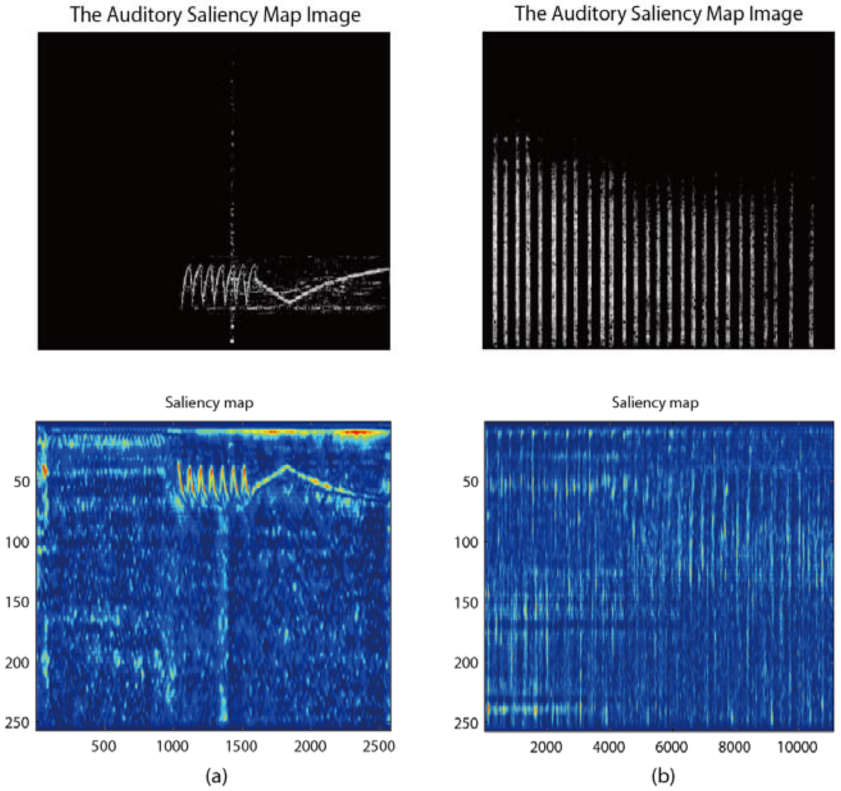


Figure 14: The final saliency detection results and the detection results by using Kayse's approach.

In the result image of the second row from Figure 13(a) which generated based on the Kayser's approach, the salient sound of track beep from example A has not been clearly located in the auditory saliency map, since its visual representation can hardly be distinguished from the background noise due to the reason of sharing a similar color. Similarly, the image of second row from Figure 13(b) shows that the multiple salient sounds of horse's hoof hitting to the ground have not been detected by the Kayser's approach, because the visual representations of these salient sounds are not clear and distinguishable. They are visually similar to the visual representation of background noise and the saliency features derived from the auditory saliency map have little corresponding relationship to the original salient sounds, which could lead to a failure in the detection of salient environmental sounds. Correspondingly, as shown in the first image of Figure 13(b), the auditory saliency map obtained from the presented approach is robust to strong background noise. It contains clear and correct saliency detection results which accurately represent the auditory saliency property of the original salient sounds in both spectral and temporal domain.

Moreover, it is also can be seen from the result images of verification test that, the accuracy and robustness of Kayser's approach will decrease sharply when the background noise is relatively strong and overlaps the salient sounds. Especially the short-term (e.g. spike-form) salient sounds can not be correctly detected when the acoustic background is composed of salient sounds as well. The explanation of this drawback to a large extent is that, the auditory saliency model of conventional approach is mainly based on the local spatiotemporal contrast and global saliency information has been rarely taken into account. Therefore, other previous mentioned auditory saliency models which based on the similar saliency detection principle could also confront the same limitation.

3.5.2. Experiments of Real Environmental Sound Recognition

3.5.2.1. Experiment Setup

The results presented in this chapter are based on the audio data that collected from the real environment which includes both indoor and outdoor environments, some of which are collected from the online sound dataset of *freesound.org* while others are recorded by using microphone. The extracted salient sounds are typical environmental sounds that exist in everyday life and can be recognized into 8 categories, which are Dog Barking (DB1), Clock Alarm (CA), Door Bell (DB2), Glass Break (GB), Phone Ring (PR), Cellphone Vibration (CV), Police Siren (PS) and Emergency Alarm (EA).

For each of these 8 classes, 6 to 10 recordings are collected with different types are included. Multiple segments of salient sounds will be obtained for experiments. All the data have a two channel tracks out of which only one channel was extracted for processing and a unified sampling rate of 44.1 kHz. Since the salient sounds are vary in time durations (0.4 sec to 10 secs), all the segments of sounds are formalized into 1 second for further processing. 75% of the sound data are used for training and the rest 25 % are applied for testing.

3.5.2.2. Experimental Protocol

In this experiment, the sound examples are processed in MATLAB software first to obtain the corresponding PSD curves. According to the mentioned way of dividing of frequency bands, all the spectral parts that larger than 20 kHz in PSD curves will not be considered in this experiment. Considering the PSD curve that larger than P_s, calculate the position of crossover point of PSD curve and the horizontal line of P_s.

For each frequency band, if there is at least one crossover point, calculate the horizontal part in the line of P_s that corresponds to the PSD value which larger than Ps. Calculate the ratio of horizontal part in each frequency band with respect to the length of frequency band and obtain the parameters of μ_1, μ_2, μ_3, μ_4 for all the experimental data.The MFCC features are calculated by using 13 linear filters and 27 log filters which lead to the 13 cepstral coefficients, 13 delta coefficients and 13 delta-delta coefficients. Particularly, the lower frequency is set to be 133.3333 Hz, the length of Hamming window is 256 points, number of FFT is 512 points and linear spacing is 66.6666. By applying the toolbox of LIBSVM, the fuzzy vectors of all the examples and the corresponding labels are used for training the SVM classifier.

3.5.2.3. Recognition Results

To evaluate the performance of proposed spectral energy based fuzzy vector feature, the classification accuracy is presented in Table 2 and the confusion matrix is constructed by applying the trained SVM model to the test data and given in Table 3.

No.	Class	FV	FV+MFCC
1	Dog Barking(DB1)	87.50	93.75
2	Clock Alarm(CA)	85.71	91.42
3	Door Bell(DB2)	83.34	91.67
4	Glass Break(GB)	93.33	97.78
5	Cellphone Vibration(CV)	90.00	94.00
6	Police Siren(PS)	96.23	98.11
7	Phone Ring(PR)	100	100
8	Emergency Alarm(EA)	85.70	90.48
	Overall Accuracy	90.90	94.65

Table 2: The Classification Accuracy (%) for SVM Classifier.

Data	Classified Sounds % (in same order as rows)							
	1)	2)	3)	4)	5)	6)	7)	8)
DB1	87.50	6.25	0	0	0	6.25	0	0
CA	0	85.71	0	0	0	0	0	14.28
DB2	8.33	0	83.34	0	0	0	8.33	0
GB	0	6.70	0	93.33	0	0	0	0
CV	0	0	0	10.00	90.00	0	0	0
PS	3.77	0	0	0	0	96.23	0	0
PR	0	0	0	0	0	0	100	0
EA	0	9.52	0	4.76	0	0	0	85.70

Table 3: Confusion Matrix for 8-Class Classification by using Fuzzy Vector Feature.

The rows of the matrix represent the sound classes that should be classified and the columns illustrate the classified results. Table 3 shows that, the classification could be mismatched when similar spectral energy distributions exist in two sounds. However, since the partition of frequency band in this thesis is empirical and it can be adjusted according to particular requirement, the fuzzy vector could be added with more atoms and more accurate results could be achieved if the partition is more sophisticated.

3.5.2.4. Discussion

The performance of proposed fuzzy vector based spectral feature of sound and the two-level classification approach is verified. Table 2 shows the classification accuracies of all the 8 class examples by using fuzzy vector (FV) and MFCC features, in which overall accuracy of 90.9% can be achieved by using the FV features. As shown in Table 2, the classification accuracies of all the 8 classes by using FV features are over 83%, while the maximum classification accuracy of 100% is achieved when classifying the Phone Ring. However, the classification accuracy of Door Bell (83.34%) is lower than others, followed by Clock Alarm (85.71%) and Emergency Alarm (85.7%), as their spectral energy are generally distributed in the lower frequency band which less than 5 kHz.

For those mismatched salient sound examples, the second-level classification process is conducted by using the 39 coefficients of MFCC based features. Table 2 shows the final accuracies of 8 class sound examples by combining MFCC based features and FV features together, in which the overall accuracy has been improved by approximately 4% and reaches 94.65% for all the test data. Meanwhile, the most significant improvement (over 8%) has been observed in the class of Door Bell (DB2). The reason for this improvement is that, the MFCC based features could represent the temporal characteristic of sound while FV features only demonstrate spectral energy distributions. Thus, if two sounds have similar FV features, they could be classified by applying the MFCC based features. Moreover, it is also shown in Table 2 that, classification accuracies of other classes have been improved respectively.

3.6. Conclusion

In this chapter, motivated by the bio-inspired saliency detection mechanism of human perception and awareness ability, a novel salient environment sound detection and recognition approach for machine awareness is presented. The proposed approach including two general parts which are auditory saliency detection and salient sound recognition, respectively. The salient environmental sound detection is based on the combination of heterogeneous saliency features: 1) acoustic saliency feature which represented by MFCC in temporal domain and the PSD of sound in spectral domain; 2) visual saliency feature from spectrogram which transform auditory saliency into image saliency. Regarding the recognition aspect of salient sound, in order to better describe the spectral differentials of different sounds, a novel spectral feature of spectral energy distribution which based on the fuzzy set theory is designed as an complementary part of MFCC based features for achieving the environmental sound recognition task.

The contributions of the presented work are as follow: 1) To provide a global saliency feature for detection and overcome the drawback of using local contrast, a short-term Shannon entropy (SSE) method is initially proposed to estimate the global saliency characteristic of environmental sound; 2) To decrease the interference of background noise along with the pseudo background noise that generated by other less salient sounds, inspired by the researches of neurobiology that inhibitory effect exists when two salient points are close to each other, a computational IOR model is proposed to verify the temporal saliency derived from MFCC; 3) To accurate detect the salient sounds in auditory saliency map, a spectral saliency detection approach is presented to obtain the saliency feature from the PSD of sound signal; 4) The image saliency feature is achieved by using the opponent color space, the goal of which is to detect the salient sounds with high power energy represented by red color region, as well as the acoustic background with green color region in a log-scale spectrogram; 5) The final saliency detection result is obtained by using a heterogeneous saliency feature fusion method; 6) To represent the spectral feature of environmental sound, a fuzzy vector based novel feature is presented by applying the fuzzy representation of spectral energy distribution

Experiment results show that, the presented saliency detection approach has a significant improvement in processing real environment sound tracks. Especially in dealing with the sounds with strong level of background noise and overlapped salient sounds, authentic salient point can be correctly detected by applying the proposed inhibition of return (IOR) model. It has also indicated that, the inhibitory effect could influence the computation of auditory saliency detection and the importance of depressing the pseudo background noise which cause by less salient sounds. The final auditory saliency detection results show the accuracy as well as the robustness of presented approach and support that my approach outperforms other conventional auditory saliency detection approaches. Meanwhile, the recognition performance of presented approach is tested by using support vector machine method and indicates that, competitive recognition accuracy can be achieved by using the combination of MFCC based acoustic features and designed fuzzy representation of spectral energy distribution based feature. The robustness as well as the effectiveness is supported by the results which show good classification accuracy in dealing with real sound signals.

Generally, the artificial awareness ability of environmental sound for machines is possible to realize by applying the presented approach. Furthermore, the obtained result could be used as the auditory awareness knowledge of surrounding environment to be fused with image knowledge for further purpose.

Chapter 4. Salient Information based Autonomous Environmental Object Detection and Classification

4.1. Introduction

Due to the importance and distinctive representation of visual information, the visual object detection and classification play a vital role in the perception process of surrounding environment in our lives. Since human beings perform this process in a joint way of using bottom-up and top-down mechanisms, bio-inspired autonomous object detection and classification approach is researched in this chapter in order to enable machines to have better visual perception ability in real environment. In this section, the introduction and motivation of my work will be illustrated in details to provide a better view for why and how I achieve this research goal.

As comprehensively discussed in previous chapter, the perception process of human beings relies on the two major mechanisms. The first one is saliency-driven bottom-up mechanism and another is task-driven top-down mechanism. Both of these two mechanisms do not work separately but in an interrelated pattern. However, due to the difference in biological priority of these two perception mechanisms, it will be more instinctive to perceive the object by using the bottom-up saliency mechanism than the other one. Nevertheless, even in the active perception process which based on the top-down mechanism, the perception process of using the bottom-up mechanism could be the very first stage which naturally performed by human brain. This probably because that, the bottom-up saliency mechanism is an unconscious selective attention pattern which will be conducted in very short times. For example, when exposed to an unfamiliar environment, we will firstly attend and analyze the salient objects which perceived by using the bottom-up saliency mechanism and to distinguish the interest foreground objects while ignoring the backgrounds. If we are looking for an expected object in the environment, the bottom-up based unconscious perception process could also be the preliminary stage in order to effectively utilize the perception resources and followed by the active searching in a top-down manner to accomplish the task.

As reviewed in Chapter 2, the traditional visual saliency detection approaches can be generally illustrated into local and global schemes. Most of them are based on the center-surround operator, contrast operator as well as some other saliency features. Since these features are mostly derived in pixel level from the original image while no intrinsic information of the object is taken into account, the detected salient regions could not cover the expected objects in certain circumstances. In (Wickens and Andre, 1990) the term of objectness is characterized as the visual representation that could be correlated with an object, thus an objectness based object shape detection approach is presented. The advantage of using objectness as the representation of object is that, the object can be considered as an integrated one for further processing based on the perceptual characteristic of our perception system, such as the human vision system. Notably, in (Alexe et., 2010) and (Alexe et., 2012) the objectness is used as a location prior to improve the object detection methods, the yielded results have shown that it outperformed other approaches, including traditional saliency, interest point detector, semantic learning and the HOG detector, and good results can be achieved in both static images and videos. Thereafter, in (Chang et al., 2011b), (Spampinato et al., 2012) and (Cheng et al., 2014) the objectness property is used as the generic cue to be combined with other saliency characteristics to achieve a better performance in salient object detection, the experimental results of which have proved that the objectness is an important property as well as an efficient way in the detection of objects and can be applied to many object-related scenarios. Therefore, it is worthy of researching the approach of detection and classification of environmental objects by using objectness and conduct it in an autonomous way.

4.2. Overview of the Approach

In general, the proposed approach in this chapter including two stages which are the detection of salient foreground objects and the classification of detected objects, respectively. To be specific, foreground objects of the environment are considered to be salient compare with the background regions because the foreground objects are much more interesting and salient in information than backgrounds to human beings in the perception of environment. The overview of the approach is shown in Fig. 14.

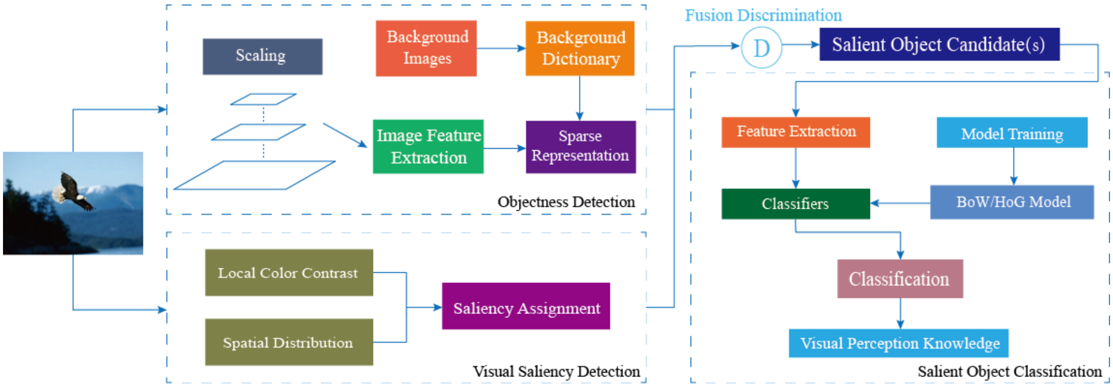


Figure 15: The general overview of proposed environmental foreground object detection and classification approach.

As demonstrated in Fig. 14, the visual image of the environment will be parallel processed by two units in the first stage, which are the objectness detection and visual saliency detection. The visual saliency detection unit is considered as the acceleration of detection in case that the object is visually salient compare to the background, thus the whole detection time could be sharply decreased. However, as common objects which exist in the environment are sometimes not visually salient, then the expected object will not be correctly perceived by the visual saliency detection unit. Therefore, the objectness detection is performed in parallel to locate the foreground objects which are salient in the term of informatics compared to the background environment. The detection results derived from these two units will be determined in order to find the expected object based on the fusion discrimination stage, while a trained classifier will be applied for classification by using proper image features.

The proposal of this framework relies on the assumption that, as inferred in the aforementioned literatures in Chapter 1, bio-inspired perception mechanism of human beings are the combination of both saliency-driven bottom-up pattern and task-driven top-down pattern. However, the bottom-up pattern is believed to be happened prior to the top-down pattern because it is a process of unconsciousness based on the saliency mechanism, while the top-down pattern will take longer time to perceive the expected object since the active searching must be carried out within the entire vision range. Obviously, foreground objects could be visually salient in some occasions while not salient in others. Therefore, the perception process based on visual saliency could be either much faster than top-down approaches when expected objects are salient, or it will completely fail if the objects are not visually salient. This shortcoming is because that, current approaches of visual saliency detection are mostly based on the contrast operation of local pixels or regions while the saliency characteristics in informatics between foreground objects and backgrounds are hardly taken into account.

Consequently, the visual saliency based object detection methods are considered as the complementary part of top-down pattern based active searching approaches in the visual perception process of my work. As shown in the objectness detection unit in Fig. 14, the foreground object is detected based on the objectness property by using the sparse representation approach. Note that the foreground object is considered to be salient from the informatics prospective, which means that the foreground objects are more interesting and contain valuable information compared to the background during the exploration of environment. By combined with the visual saliency information of object, the detection process could be accelerated in certain circumstances when the expected foreground object is visually salient, while the detection accuracy could be maintained without significant decrease when the foreground object is not visually salient but informatics salient compared to the background.

In addition, the fusion discrimination unit is used to determine whether the salient region or the detect windows which contain most objectness should be selected as the salient object candidates. Afterwards, these candidates will be classified by using the pre-trained classifiers to obtain the visual information of foreground objects, which yields the final knowledge of visual perception for further processing. Therefore, the detection of salient foreground environmental object can be performed by machines in an autonomous way and preliminarily similar to the way that how human beings achieve the environment perception goal.

4.3. Sparse Representation based Salient Environmental Object Detection

4.3.1. Image Feature Extraction

4.3.1.1. Gabor Filter

The Gabor filter was initially proposed in (Gabor, 1946) which goal is to analyze the time-frequency characteristic of one-dimensional signal. It has been widely known that the classic Fourier transform has high frequency resolution but no time resolution because the basis functions of Fourier transform are infinite along time axis. Thus, the signal could be analyzed within a local interval by using a Gaussian window function to provide the time-frequency resolution.

To be specific, the one-dimensional Gabor filter can be defined as multiplication of the cosine or sine function with a Gaussian window as follows

$$g_c(x) = \frac{1}{\sqrt{2\pi\sigma_g}} e^{-\frac{x^2}{2\sigma_g{}^2}} \cos(2\pi\omega_0 x) \quad \textbf{(21)}$$

$$g_s(x) = \frac{1}{\sqrt{2\pi\sigma_g}} e^{-\frac{x^2}{2\sigma_g{}^2}} \sin(2\pi\omega_0 x) \quad \textbf{(22)}$$

in which the Gabor filter has maximum response at frequency of ω_0 and σ_g is the spread of each Gaussian window.

Particularly, the Gabor function can be seen as a linear filter while the response of Gabor function in frequency and orientation is similar to the biological representation of simple cells in human visual system. Therefore, (Marčelja, 1980) and (Daugman, 1980) have suggested that the characteristics of simple cells in the visual cortex of mammalian brains can be modeled by Gabor functions, thus in (Daugman, 1985) the Gabor filter was extended to two-dimensions which can be defined as

$$g_c(x,y)=\frac{1}{\sqrt{2\pi\sigma_x\sigma_y}}e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x}+\frac{y^2}{\sigma_y}\right)}\cos\left(2\pi\omega_{x_0}x+2\pi\omega_{y_0}y\right) \qquad \textbf{(23)}$$

$$g_s(x,y)=\frac{1}{\sqrt{2\pi\sigma_x\sigma_y}}e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x}+\frac{y^2}{\sigma_y}\right)}\sin\left(2\pi\omega_{x_0}x+2\pi\omega_{y_0}y\right) \qquad \textbf{(24)}$$

where $\left(\omega_{x_0},\omega_{y_0}\right)$ are the centre frequency of x and y axis, $\left(g_c,g_s\right)$ are the Gabor filters of even (e.g. cosine) and odd (e.g. sine) waves while $\left(\sigma_x,\sigma_y\right)$ denotes the length of Gaussian window.

(Henriksen, 2007) argued that, in the application of image processing, the impulse response of Gabor filter is defined as the multiplication of a sinusoidal wave and the Gaussian function that has a real and an imaginary component representing orthogonal directions. Both of these two components can be formed into a uniformed complex version or be used individually as

$$g_{complex}\left(x,y;\lambda,\theta_g,\psi,\sigma_{sg},\gamma\right)=\exp\left(-\frac{x^{'2}+\gamma^2y^{'2}}{2\sigma_{sg}^{\ 2}}\right)\exp\left(i\left(2\pi\frac{x^{'}}{\lambda}+\psi\right)\right) \qquad \textbf{(25)}$$

$$g_{real}\left(x,y;\lambda,\theta_g,\psi,\sigma_{sg},\gamma\right)=\exp\left(-\frac{x^{'2}+\gamma^2y^{'2}}{2\sigma_{sg}^{\ 2}}\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right) \qquad \textbf{(26)}$$

$$g_{imaginary}\left(x,y;\lambda,\theta_g,\psi,\sigma_{sg},\gamma\right)=\exp\left(-\frac{x^{'2}+\gamma^2y^{'2}}{2\sigma_{sg}^{\ 2}}\right)\sin\left(2\pi\frac{x'}{\lambda}+\psi\right) \qquad \textbf{(27)}$$

where $x^{'}=x\cos\theta_g+\text{y}\sin\theta_g$ and $y^{'}=-x\sin\theta_g+\text{y}\cos\theta_g$. In the above equations of Eq. (25) to Eq. (27), the λ is the wavelength of the sinusoidal wave, θ_g represents the orientation of the parallel stripes of each Gabor function, ψ is the phase offset, σ_{sg} is the standard deviation of Gabor function and the γ is the spatial aspect ratio. It is obviously seen that the Gabor filters can be adjusted to yield a better resolution by applying different parameters.

4.3.1.2. 2-D Gabor Feature Extraction

Since the kernel of Gabor filters is believed to be a good simulation model which similar to the receptive field profiles of cortical simple cells (Field, 1994), Gabor filter is widely used to capture the local characteristic of image in multiple frequencies (scales) and orientations due to the good performance of spatial localization as well as orientation selection. The two-dimensional Gabor function can therefore enhance the features of edge, peak and ridge and robust to illumination and posture to a certain extent.

Considering the statistic property of image, the kernel of Gabor function can be defined as (Liu and Wechsler, 2002)

$$\psi_{u,v}\left(x,y\right)=\frac{\left\|k_{u,v}\right\|^{2}}{\sigma_{gr}{}^{2}}\exp\left(-\frac{\left\|k_{u,v}\right\|^{2}\left(x_{l}^{2}+y_{l}^{2}\right)}{2\sigma_{gr}{}^{2}}\right)\cdot(\exp(\mathrm{i}\,k_{u,v}\cdot\begin{pmatrix}x_{l}\\ y_{l}\end{pmatrix})-\exp(-\frac{\sigma_{gr}{}^{2}}{2})) \quad \textbf{(28)}$$

where *u* and *v* represent the orientation and scale of the Gabor kernels, x_l and y_l are the coordinates of pixel location, ||·|| denotes the norm operator and σ_{gr} determines the ratio of the Gaussian window width to wavelength. Particularly, the wave vector $k_{u,v}$ is defined as follows

$$k_{u,\mathrm{v}}=k_{\mathrm{v}}e^{i\phi_{u}} \quad \textbf{(29)}$$

where $k_{\mathrm{v}}=k_{\max}/f_{sf}{}^{\mathrm{v}}$ and $\phi_{u}=\pi u/8$, in which $k_{\max}$ is the maximum frequency and f_{sf} is the spacing factor between kernels in the frequency domain (Lades et al., 1993). By using different values of *u* and *v*, a set of Gabor filters with different scales and orientations can be obtained.

In addition, the Gabor feature of an image is the convolution of the image with a set of Gabor filters in the filter bank which defined by Eq. (28). The formulation of the Gabor feature derived from the image *I*(*z*) in this chapter can be defined as (Yang and Zhang, 2010)

$$G_{u,v}\left(z\right)=I\left(z\right)*\psi_{u,v}\left(z\right) \tag{30}$$

where $z=(x_l,\ y_l)$ denotes the location of a pixel, $G_{u,v}(z)$ is the Gabor feature of an image $I(z)$ in orientation u and scale v, the * represents the convolution operator. According to the convolution theorem, the $G_{u,v}(z)$ from Eq. (30) can be also derived by using Fast Fourier transform (FFT) as

$$\mathrm{F}\ \left\{G_{u,v}\left(z\right)\right\}=\mathrm{F}\ \left\{I\left(z\right)\right\}\mathrm{F}\ \left\{\psi_{u,v}\left(z\right)\right\} \tag{31}$$

$$G_{u,v}\left(z\right)=\mathrm{F}^{\ -1}\left\{\mathrm{F}\ \left\{I\left(z\right)\right\}\mathrm{F}\ \left\{\psi_{u,v}\left(z\right)\right\}\right\} \tag{32}$$

where F and $\mathrm{F}^{\ -1}$ represent the Fourier and inverse Fourier transform. Note that the $G_{u,v}(z)$ is a complex number that could be written in the form as

$$G_{u,v}\left(z\right)=M_{u,v}\left(z\right)\cdot\exp\left(i\theta_{u,v}\left(z\right)\right) \tag{33}$$

in which the $M_{u,v}(z)$ is the magnitude and $\theta_{u,v}(z)$ is the phase. Meanwhile, as discussed in (Jones and Palmer, 1987) and (Burr et al., 1989), frequently used orientations are $u\in\left\{0,\ldots,7\right\}$ and $v\in\left\{0,\ldots,4\right\}$.

Regarding the practical requirement of my work, the objects in environment are regular in shape and contour, thus the scale and orientation parameters are set to be 3 and 2, respectively. In Fig. 15, the structure of one Gabor filter is graphically demonstrated as an example.

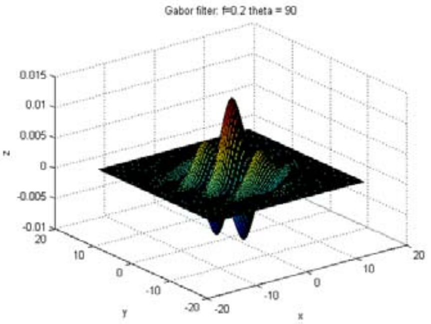


Figure 16: The spatial structure of a Gabor filter.

In addition, the spatial representations of the Gabor filters are presented in Fig. 16. The orientations of obtained Gabor filters including both horizontal and vertical directions, while the scale parameter is set to be 3 empirically to make compromise to computation cost as well as the accuracy in this thesis.

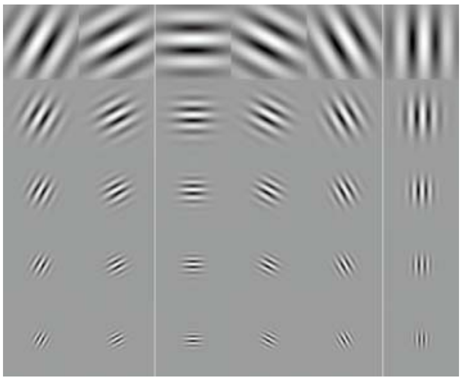


Figure 17: Representation of two-dimensional Gabor filters in spatial domain.

4.3.2. Visual Saliency Detection

In real application scenarios, the images of environment that collected by sensors are commonly in the RGB color space to provide vivid representations for human vision. Therefore, the saliency property of local color contrast should be taken into account in order to better describe the rarity characteristics of objects. Meanwhile, as the saliency property of objects correlate to the global contrast between background and foreground, the global saliency features including distributions of colors should be also considered as important information in visual saliency detection to distinguish the objects. Instead of designing the visual saliency algorithm, I applied a state of the art saliency detection approach proposed in (Perazzi et al., 2012). The approach is based on the saliency property of color contrast along with its spatial distribution and has been proved to be effective as well as fast in practical applications.

The general proposal is based on the combination of two well-defined contrast measures including uniqueness and spatial distribution of homogeneous elements. The image will be decomposed into basic elements that preserve relevant structure, each of which is generated by clustering pixels with similar color into perceptual regions. Therefore, the boundaries of these elements will be preserved as edges are considered to be interesting while the localized elements could provide compactness to visual perception constrained to shape and size. To be specific, the image is abstracted based on an adaptation of SLIC superpixels approach which proposed in (Achanta et al., 2010). Instead of using the K-means clustering method in RGBXY space, the adapted SLIC approach uses K-means clustering in geodesic image distance which proposed by (Criminisi et al., 2010) in CIELab space. Compare to the original RGBXY space, the geodesic image distance is able to achieve the objective of connectivity while the advantages of SLIC superpixels can be retained. Notably, the presented abstraction process can be seen as an approximative way of detecting objectness, in which the pixels with similar color could indicate the same object and thus the clustered regions can be seen the representations of objects with different objectness. In general, the saliency feature is calculated by measuring two kinds of local contrast features, which are element uniqueness and element distribution, respectively. However, only the brief introduction will be presented in this section to illustrate the calculation procedure as the detailed information of this approach could be found in (Perazzi et al., 2012).

The element uniqueness is defined as

$$U_i = \sum_{j=1}^{N} \left\| c_i - c_j \right\|^2 \cdot \omega\left(p_i, p_j\right) \qquad \textbf{(34)}$$

where c_i, c_j are the CIELab colors of segment i and j, p_i and p_j denote the positions of segment i and j, respectively, $\omega(p_i, p_j)$ is the weight relates to the distance between segments and can be written as $\omega_{ij}^{(p)}$ which determines the form of the uniqueness operator. To be specific, the local contrast similar to the center-surround operator is obtained when the weight $\omega(p_i, p_j)$ is a local function, which means that the weight of large distance will be assigned small. On the other hand, a global contrast operator can be achieved when $\omega(p_i, p_j)$ is a constant to perform the contrast operation between regions. Particularly, the global saliency features proposed in (Achanta et al., 2009) and (Cheng et al., 2011) are the approximations of Eq. (34) when $\omega_{ij}^{(p)}$=1. Moreover, Eq. 4-14 can be decomposed by factoring out the quadratic error function as follows

$$U_i = c_i^2 \sum_{j=1}^{N} \omega_{ij}^{(p)} - 2c_i \sum_{j=1}^{N} c_j \omega_{ij}^{(p)} + \sum_{j=1}^{N} c_j^2 \omega_{ij}^{(p)} \tag{35}$$

in which the weight is set to be a Gaussian kernel as $\omega_{ij}^{(p)} = \frac{1}{Z_i} \exp\left(-\frac{1}{2\sigma_p^2} \left\| p_i - p_j \right\|^2 \right)$, σ_p determines the operation range and Z_i is a normalization factor that ensures the sum result of the first term of Eq. (35) equals 1, the $\sum_{j=1}^{N} c_j \omega_{ij}^{(p)}$ and $\sum_{j=1}^{N} c_j^2 \omega_{ij}^{(p)}$ can be treated as two filters which perform filtering operation on color c_j and squared color c_j^2 along *x* and *y* axis of the image. Consequently, the Eq. (34) could be evaluated in linear time to decrease the computation time.

The spatial distribution of element is used to measure the color uniqueness over spatial structure of the image that can be defined as

$$D_i = \sum_{j=1}^{N} \left\| p_{mj} - \mu_i \right\|^2 \underbrace{\omega\left(c_i, c_j\right)}_{\omega_{ij}^{(c)}} \tag{36}$$

where $\omega_{ij}^{(c)}$ describes the color similarity of segments *i* and *j*, $\mu_i = \sum_{j=1}^{N} \omega_{ij}^{(c)} p_j$ denotes the weighted mean position of color c$_i$. The spatial variance of D_i in Eq. (36) could be used as the indication of spatially compact objects, indicating that elements with low variance should be considered as salient objects compared with the elements that spatially widely distributed. Thereafter, the saliency value for each element can be calculated as

$$Sv_i = U_i \cdot \exp\left(-k_s \cdot D_i\right) \tag{37}$$

in which the exponential function $\exp\left(-k_s \cdot D_i\right)$ is applied to emphasize D_i in order to maintain the higher significance and discriminative power and the scaling factor k_s is set to be 6 experimentally in Perazzi's work. Thus, the final saliency value of each pixel in image is defined as a weighted linear combination as

$$\hat{S}v_i = \sum_{j=1}^{N} \omega_{ij} Sv_j \tag{38}$$

in which the weight is based on a Gaussian kernel given as

$$\frac{1}{Z_i}\exp\left(-\frac{1}{2}\left(\alpha_v\left\|c_i-c_j\right\|^2+\beta_v\left\|p_i-p_j\right\|^2\right)\right) \qquad \textbf{(39)}$$

where the α_v and β_v are parameters controlling the sensitivity to color and position.

4.3.3. Sparse Representation based Foreground Objectness Detection

4.3.3.1. Motivation

Compare to the environmental background, the foreground objects contain more useful and semantic information in the perceptual process from the perspective of human visual. It makes the foreground objects more interesting and valuable that can be treated as salient in informatics. Therefore, the detection of foreground objects is considered to be a crucial and fundamental task in the perception of environment for artificial machines.

From visual saliency point of view, the foreground objects could be either salient to human vision or non-salient in some scenarios. Inspired by the observation derived from the perception mechanism of human visual system, a unified perception process including both saliency-driven bottom-up pattern and goal-driven top-down pattern is considered to be the simulation of human perception for artificial awareness. Practical speaking, although various visual saliency approaches have been proposed and could be used to extract the visually salient objects in a highly efficient bottom-up manner, the active searching of visually non-salient objects should be also taken into account in a top-down manner and combined to improve the performance of detection. Meanwhile, the term of objectness provides a better description of object which can be used in the many applications. In (Ali Shah et al., 2013), (Zhang and Zhang, 2014), (Zhang and Liu, 2014) and (Cheng et al., 2014), the objectness measure has been used as an important feature of object to improve the performance of the original methods in several applications, such as automatic object detection and salient region detection in images. The experimental results have shown the efficiency and effectiveness in improving the detection accuracy and decreasing the computational cost, which reveal the importance of using the objectness feature for object detection.

Moreover, inspired by the early works of (Olshausen, 1996) and (Olshausen and Field, 1997) which revealed the biological foundation of sparse coding, researches of (Elad and Aharon, 2006b), (Mairal et al., 2008) and (Wright et al., 2009) have shown that the sparse representation is considered to be an extremely powerful mathematical tool for representing and compressing high-dimensional signals in many computer vision applications, including natural image restoration, image denoising and human face recognition. In the work of (Yang et al., 2009), a novel spatial pyramid matching (SPM) approach is proposed for image classification. Instead of using the traditional vector quantization (VQ) technique, selective sparse coding of SIFT feature matrix is applied to extract salient properties of local image patches. The author suggested that the sparse codes of image features such as SIFT descriptor could be more useful than other traditional techniques in the field of image processing. Recently, (Zhang et al., 2013) proposed a novel image classification approach based on spatial pyramid robust sparse coding as an improvement of (Yang et al., 2009). The proposed method applied robust sparse coding in both of the construction of codebook and the coding of local image features, in which the spatial constraint is combined to efficiently encoding the local feature in order to improve the robustness of traditional BoW model. In general, all these works have shown the effectiveness of sparse representation in encoding the image features. However, the natural characteristics of objects themselves should be also taken into account to improve the performance.

As motivated before, a reconstruction error based foreground object detection approach by using the sparse representation is proposed in this section. Different from other approaches, the objectness characteristic of potential object in this method is obtained by calculating the reconstruction error of the object feature matrix over an overcomplete background dictionary which describes the dissimilarity between object and background. Since the theoretical basis and derivation of sparse representation has been well studied, the detailed introduction of sparse coding is omitted while the illustrations of key components of my approach will be given in this section.

4.3.3.2. Background Dictionary Learning

According to the basic model of sparse representation, the sparse representation of a column signal $x \in \Re^n$ with respect to an overcomplete dictionary $D \in \Re^{n \times K}$ including *K* atoms can be described by the following sparse approximation problem as

$$\min_{\alpha} \|\alpha\|_0 \quad \text{subject to } \|x - D\alpha\|_2 \leq \varepsilon \qquad \textbf{(40)}$$

where $\|\cdot\|_0$ is the l^0-norm which counts the nonzero entries of a vector, α is the sparse coefficient and ε is the error tolerance. According to the discussion in the work of (Davis et al., 1997), extract determination of sparsest representation which defined in Eq. (40) has been known as a non-deterministic polynomial (NP) -hard problem. This means that the sparsest solution of Eq. (40) has no optimal result but trying all subsets of the entries for *x* which could be computational unavailable. (Donoho, 2006) and (Candes et al., 2006) have argued that if the sought solution *x* is sparse enough, the solution of the l^0-norm problem could be replaced by the approximated version of the l^1-norm as

$$\min_{\alpha} \|\alpha\|_1 \quad \text{subject to } \|x - D\alpha\|_2 \leq \varepsilon \qquad \textbf{(41)}$$

where $\|\cdot\|_1$ is the l^1-norm. The similarity in finding sparse solution between using the l^1-norm and the l^0-norm has been supported by the work of (Donoho and Tsaig, 2008).

In the real application of sparse representation, the performance of each approach relies on the choice of the overcomplete dictionary *D*. Current dictionary learning methods can be categorized into two kinds based on the discussion in (Rubinstein et al., 2010), which are the analytic approach and the learning-based approach. The first approach refers to the dictionaries which generated from the standard mathematical models, such as Fourier, Wavelet and Gabor, to name a few, which have no semantic meaning correlate to the natural images. On the other hand, the second approach uses machine learning based techniques to generate the dictionary from image examples. Therefore, the obtained dictionary could represent the examples in a close manner. Compared to the first approach which prespecifies the dictionary atoms, the second approach is an adaptation process between dictionary and examples from the machine learning perspective. Though the analytic dictionary is simple to be implemented and highly structured, the learning-based dictionary has shown a better performance in the application of image processing. Regarding the requirement of my work, the dictionary that learned based on image examples is considered to provide a better and semantic description of the background.

As discussed in many dictionary learning methods (see (Aharon et al, 2006) and (Lee et al., 2006) as examples), the optimization problem is convex in dictionary *D* while coefficients *α* is fixed and vice versa, but not convex in both simultaneously, the iteration operation of this calculation is considered in most of the dictionary learning algorithms. Therefore, the learning of the dictionary can be generally illustrated in three steps in order to find the sparse solution of optimal dictionary *D* and sparse coefficients of *α*. First, considering a general problem defined by formulation in the form of *F*-norm as

$$\min_{D,\alpha} \left\| x - D\alpha \right\|_F^2 + \gamma_c \left\| \alpha \right\|_1 \text{ subject to } \forall i, \left\| \alpha_i \right\|_0 \leq c \qquad \textbf{(42)}$$

where *γ_c* is a constant. Thereafter, initialize the dictionary *D* randomly. The second step is to obtain the sparse coefficients while make the dictionary fixed as

$$i = 1, 2, \ldots, N, \quad \min_{\alpha_i} \left\| x_i - D\alpha_i \right\|_F^2 \text{ subject to } \forall i, \left\| \alpha_i \right\|_0 \leq c \qquad \textbf{(43)}$$

When the sparse coefficients *α* have been obtained, the third step is to solve the formulation in Eq. (42) to generate the optimal bases *D* while the coefficients *α* is fixed as

$$\min \left\| x - D\alpha \right\|_F^2 \text{ subject to } \sum\nolimits_{i=1}^{k} D_{i,j}^2 \leq c, \forall j = 1, \ldots, n. \qquad \textbf{(44)}$$

where $D_{i,j}$ denotes the element of i^{th} row and j^{th} column in the dictionary matrix D. The above Eq. (44) is a least squares problem with quadratic constraints that can be solved by using a Lagrange dual. Considering the Lagrangian as

$$\mathrm{L}\left(\mathrm{D},l\right)=trace\left(\left(x-D\alpha\right)^{\mathrm{T}}\left(x-D\alpha\right)\right)+\sum_{j=1}^{n}l_j\left(\sum_{i=1}^{k}D_{i,j}^{2}-c\right) \quad \textbf{(45)}$$

where $l_j \geq 0$ is a dual variable. By minimizing over dictionary D, the Lagrange dual is obtained as

$$D\left(l\right)=\min_{D} L\left(D,l\right)=trace\left(x^{\mathrm{T}}x-x\alpha^{\mathrm{T}}\left(\alpha\alpha^{\mathrm{T}}+\Lambda\right)^{-1}\left(x\alpha^{\mathrm{T}}\right)^{\mathrm{T}}\right)-c\Lambda \quad \textbf{(46)}$$

where $\Lambda = diag\left(l\right)$. The gradient and Hessian of $D\left(l\right)$ are calculated as follows

$$\frac{\partial D\left(l\right)}{\partial l_i}=\left\|x\alpha^{\mathrm{T}}\left(\alpha\alpha^{\mathrm{T}}+\Lambda\right)^{-1}e_i\right\|^{2}-c \quad \textbf{(47)}$$

$$\frac{\partial^{2} D\left(l\right)}{\partial l_i \partial l_j}=-2\left(\left(\alpha\alpha^{\mathrm{T}}+\Lambda\right)^{-1}\left(x\alpha^{\mathrm{T}}\right)^{\mathrm{T}}x\alpha^{\mathrm{T}}\left(\alpha\alpha^{\mathrm{T}}+\Lambda\right)^{-1}\right)_{i,j}\left(\left(\alpha\alpha^{\mathrm{T}}+\Lambda\right)^{-1}\right)_{i,j} \quad \textbf{(48)}$$

where $e_i \in \Re^{n}$ is the i-th unit vector. Therefore, the optimal dictionary D is given by maximizing $\mathrm{D}\left(l\right)$ as

$$D^{\mathrm{T}}=\left(\alpha\alpha^{\mathrm{T}}+\Lambda\right)^{-1}\left(x\alpha^{\mathrm{T}}\right)^{\mathrm{T}} \quad \textbf{(49)}$$

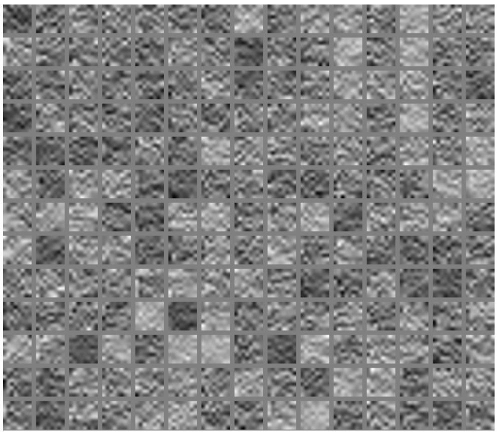


Figure 18: The learned background dictionary.

Accordingly, the dictionary applied in this chapter is derived from the 2-D Gabor feature matrix which based on the previous section of gabor feature extraction. Thus, the dictionary *D* is used to represent the image features of backgrounds. By using the dictionary, sparse coding is able to approximately represent the input features as a linear combination of a few sparse atoms. The learned background dictionary is shown in Figure 17.

4.3.3.3. Foreground Object Detection based on Representation Error

When the overcomplete dictionary D of background is learned, the objectness property of foreground object can be obtained by calculating the reconstruction error of input feature vector derived from a detection window over the learned dictionary. The underlying assumption of this approach is that, as the representation of a local image patch, each local feature vector contains the objectness property of a detection window which can be characterized as the dissimilarity between input feature vector and background dictionary. By using the sparse representation coefficients *α* of a feature vector generated from the dictionary, the reconstructed feature vector could be restored by applying an inverse operation of sparse decomposition. However, since the reconstructed feature vector derived from sparse coding is the approximation of the original feature vector, a reconstruction error between these two vectors can be calculated to indicate the dissimilarity between the current local image patch and the background image. Thus, the objectness property of each detection window could be measured for foreground object detection.

Generally, assume x_i, *i=1,…,N* is the corresponding feature vector for i^{th} local image patch, the sparse coefficient can be computed by coding each x_i over the learned dictionary D based on the l^1-minimization as

$$\min_{\alpha} \|\alpha\|_1 \text{ subject to } x = D\alpha \qquad \textbf{(50)}$$

To obtain the sparse coefficient *α*, various decomposition approaches have been proposed and proved to be effective, such as Basis Pursuit (BP) in (Chen and Donoho, 1998), Matching Pursuit (MP) in (Mallat and Zhang, 1993), Orthogonal Matching Pursuit (OMP) in (Pati et al., 1993) as well as Least Absolute Shrinkage and Selection Operator (LASSO) in (Tibshirani, 1996). Considering the computational cost and the requirement of research goal, the LASSO algorithm is applied to compute the sparse coefficients α of the input feature vector.

Thus, the reconstructed feature vector $\hat{x}$ can be calculated based on the sparse coefficients as

$$\hat{x} = D\alpha \tag{51}$$

Since $\hat{x}$ is only the approximation solution of the original x, the reconstruction error can be quantitatively given as

$$\varepsilon = \|x - \hat{x}\|_2^2 = \|x - D\alpha\|_2^2 \tag{52}$$

where $\|\cdot\|_2^2$ denotes the Euclidean distance.

Particularly, as the input image is processed in multiple scales in my research to reveal the characteristics of objects in different sizes, the input feature vector x_i of each scale will be evaluated differently as

$$\varepsilon_{fo} = \sum\left\{\forall \varepsilon_i^{s_j} > \rho_{s_j}\right\}, \quad j = 1,2,3 \tag{53}$$

where *i* denotes the reconstruction error of *i*th local image patch in each scale and ε_{fo} represents the set consists of reconstruction errors $\varepsilon_i^{s_j}$ that larger than the error threshold of ρ_{s_j} in s_j scale. Therefore, the objectness of semantic salient object can be extracted by finding the detection window which indicated by ε_{fo}.

4.3.4. Fusion Discrimination

To achieve a human-like perception process, the aforementioned visual saliency detection approach is combined to accelerate the computation time when the expected object is actually visually salient. When the expected object is not visually salient that could make the detection process based on visual saliency failed, the classical active searching of expected foreground objects based on the proposed objectness detection approach is performed in order to maintain the detection accuracy.

The fusion discrimination of these two processing procedure can be given as

$$C_{object} = \begin{cases} O_{classify}\left\{f_{vs}\right\}, & if \quad O_{classify}\left\{f_{vs}\right\} \notin \varnothing \\ O_{classify}\left\{f_{objectness}\right\} & if \quad O_{classify}\left\{f_{vs}\right\} \in \varnothing \end{cases} \tag{54}$$

In Eq. (54), C_{object} represents the classification results of foreground objects, $O_{classify}\left\{\cdot\right\}$ denotes the object classification operator, f_{vs} and $f_{objectness}$ represent the features derived from the visual saliency algorithm and objectness extraction algorithm, respectively. The general objective of Eq. (54) is to achieve a simulated classification process similar to the perception procedure of human beings, in which the visual saliency guides the perception in a prior and preliminary position based on the bottom-up mechanism. When the expected object is not visually salient, the active searching of object is performed to accomplish the perception task in a top-down manner. Therefore, the fusion discrimination allows the proposed foreground object detection approach to either perceive the salient object or find the expected object for further processing.

4.4. Salient Foreground Environmental Object Classification

4.4.1. General Introduction

When the local image patches with high objectness characteristics compared with backgrounds have been successfully detected and extracted from the original image, they are considered as the candidates of foreground objects which are interesting for perception and salient in informatics, while some of which could be visually salient as well. Therefore, the classification of the foreground object candidates is performed in order to accurately classify whether these regions contain expected objects or not.

As illustrated in Chapter 2, many popular classification models have been proposed for object classification and proved to be effective in most of the application scenarios. Considering the characteristics of typical environmental objects that can generate sounds, most of them could be similar to each other in shape, contour and sometimes in color. For example, in the indoor environment, different kinds of clock could share the same shape in order to be easily recognized. While in the outdoor environment, the fire trucks are generally the same in both shape and color. Thus, the bag-of-visual-word (BoW) model is considered in my approach as the representation model of environmental object classification.

The BoW model could be seen as a vector quantization operation (Gray, 1984) that mapping a high-dimensional feature vector to a low-dimensional vector, in which a codebook is obtained to provide the nearest neighbor (NN) representation for the original feature vector. Assume the input feature vector is $x_v \in \Re^m$, then a codebook can be given as

$$M\left(x_v\right) \in C = \left\{C_i \mid C_i \in \Re^m, i \in \left\{1,\ldots,k\right\}\right\} \tag{55}$$

where C is the set of codebook, C_i is the codeword and $M(\cdot)$ is the mapping operation. Note that the size of codebook C is k. Generally, the codeword is generated by using the K-means clustering on all the input feature vectors and can be referred as the K-means centers. By calculating the histogram of feature vector over the codebook, each vector will be represented by the assigned codeword.

(Brandt, 2010) argued that the quality of the obtained codebook can be measured as the averaged distortion between feature vector x_v and its mapping operation $M(\cdot)$ as

$$D\left(M\right) = \left\|d\left(x_v, M\left(x_v\right)\right)\right\|_E \tag{56}$$

where $d(\cdot)$ is calculation operator of distortion, $\left\|\cdot\right\|_E$ is the Euclidean distance as an typical form of measurement (Jégou et al., 2011). Accordingly, the triangle inequality form can be obtained based on the Eq. (56) as

$$\left\|d\left(x_v, x_v'\right) - d\left(x_v, M\left(x_v'\right)\right)\right\|_E \le D\left(M\right) \tag{57}$$

which suggests that the codebook could be effectiveness when one sample in a pair is approximated by its quantization result and an upper bound on the expected error for estimating the inter-sample distances is exist (Wang et al., 2014).

To ensure the codebook is optimal, two requirements should be satisfied during the clustering operation of *K*-means as

$$M\left(x\right)=\left\{c_i \mid d\left(x_v,C_i\right)\le d\left(x_v,C_j\right),\forall j\in\left\{1,\ldots,k\right\}\right\} \qquad \textbf{(58)}$$

$$C_i=\arg\min_{x'}\left\|d\left(x_v,x_v'\right)\mid x_v,x_v'\in Q_i\right\|_E \qquad \textbf{(59)}$$

where $Q_i=\left\{x_v \mid x_v\in\Re^m, M\left(x\right)=C_i\right\}$ denotes the unique Voronoi cell that describes the mapping of each set of vectors to the same codeword C_i (Wang et al., 2014). The first requirement makes the samples in each cell close to its centroid while the second requirement suggests that the codeword for a given cell must be the expectation of the data points within the cell (Lloyd, 1982).

4.4.2. Object Feature Extraction

The classification is based on the image feature extracted from all the detection windows which have salient objectness properties to determine the specific categories of potential object candidates. As mentioned before, the SIFT feature is proved to be a state of the art image feature in object detection and classification. However, due to the large computational cost and the core detection of local interest points, the SIFT feature could face an inevitable failure in dealing with objects with circular shape and smooth edge. Since environmental objects could contain various kinds of objects with different shapes, many of which are in circular shapes and have smooth edges because of the photographing condition, a variantion of SIFT feature which is the dense SIFT feature (Bosch et al., 2007) is applied in my work.

Different from the original SIFT feature, dense SIFT (DSIFT) computes the local image feature on a regular grid in just one single scale instead of multi-scale Gaussian space applied in SIFT feature computation. At each point of the dense grid, a dense set of multi-scale SIFT descriptor is efficiently computed over four circular support patches with different radii. Therefore, four descriptors of each position of dense grid could be obtained as the representation of local pixels, while multiple descriptors are computed to allow for variation in scale between images. Note that DSIFT feature is similar to the previous mentioned HoG feature since they both characterize edginess as well as orientation around pixels, however they are different in the specific way of computation. Moreover, though the DSIFT itself does not require scaling operation as it can be seen as a densely sampled version of SIFT feature extraction, the original image need to be scaled in order to better reveal the multi-scale characteristics of visual objects.

In my work, the DSIFT features of local image patches are extracted on three scales which are 16×16 pixel, 32×32 pixel and 64×64 pixel while using a detection window of 8×8 pixel. The obtained features of all the detection windows are concatenated to form a 128-dimensional vector which applied in further classification model.

4.4.3. Model Training

The classification of environmental objects is considered to be a multi-class classification problem which goal is to accurately recognize the expected object. The popular support vector machines (SVM) (Cortes and Vapnik, 1995) is used to train the classifier for environmental object classification based on LIBSVM (Chang and Lin, 2011a). Since the classic SVM is a two-class classifier, extension should be made in order to be applied in the multi-class scenario. Based on the discussion in (Hsu and Lin, 2002), the multi-class classification approaches could be described in two general categories which are one-against-all method and one-against-one method (Knerr et al., 1990), respectively.

In this sub-section, the introduction of SVM is given in a general way since I will directly use the state-of-the-art SVM toolbox for training. To be specific, there are *k* SVM models for *k* classes of objects in one-against-all method in which each SVM model is trained with all the examples in corresponding class with positive labels, while other examples are considered to be negative labels.

Let $\{(x_i, y_i) \mid x_i \in R^n, i = 1, \dots, l_s\}$ be the training examples with number of l_s in which x_i is the data while y_i is the class of x_i, then the i^{th} SVM model solves the problem defined as

$$\begin{aligned} &\min_{\omega^i, b^i, \xi^i} \quad \frac{1}{2}\left(\omega^i\right)^{\mathrm{T}} \omega^i + C_s \sum_{j=1}^{l_s} \xi_j^i \\ &\text{subject to} \quad \left(\omega^i\right)\phi\left(x_j\right) + b^i \geq 1 - \xi_j^i, \text{ if } y_i = i, \\ &\qquad \left(\omega^i\right)^{\mathrm{T}} \phi\left(x_j\right) + b^i \leq -1 + \xi_j^i, \text{ if } y_i \neq i, \\ &\qquad \xi_j^i \geq 0, j = 1, \dots, l_s, \end{aligned} \tag{60}$$

where x_i are mapped to a higher dimensional space by the mapping function □ and C_s is a positive regularization parameter.

For one-against-one method, there are *k*(*k-1*)*/2* classifiers to be trained based on the data from only two classes. Therefore, for any two classes of i^{th} and j^{th}, the problem defined by Eq. (60) is then redefined as

$$\begin{aligned} &\min_{\omega^{ij}, b^{ij}, \xi^{ij}} \quad \frac{1}{2}\left(\omega^{ij}\right)^{\mathrm{T}} \omega^{ij} + C \sum_{p} \xi_p^{ij} \\ &\text{subject to} \quad \left(\omega^{ij}\right)\phi\left(x_p\right) + b^{ij} \geq 1 - \xi_p^{ij}, \text{ if } y_p = i, \\ &\qquad \left(\omega^{ij}\right)^{\mathrm{T}} \phi\left(x_p\right) + b^{ij} \leq -1 + \xi_p^{ij}, \text{ if } y_i = j, \\ &\qquad \xi_p^{ij} \geq 0, \end{aligned} \tag{61}$$

By the definition, the solutions of these two methods are different as well. For the one-against-all method, l_s decision functions can be given by solving the Eq. (60) as

$$\begin{gathered} \left(\omega^1\right)^{\mathrm{T}} \phi(x) + b^1, \\ \vdots \\ \left(\omega^{l_s}\right)^{\mathrm{T}} \phi(x) + b^{l_s}. \end{gathered} \tag{62}$$

Thus, the example x is in the class with largest value of the decision function which can be given as

$$c_x \equiv \arg\max_{i=1,\ldots,l_s}\left(\left(\omega^i\right)^{\mathrm{T}}\phi(x)+b^i\right) \qquad \mathbf{(63)}$$

While on the other hand, the one-against-one method defined in Eq. (61) can be solved by using the voting strategy (Friedman, 1996), in which the final classification results of all the examples are based on the highest votes corresponding to a class.

Particularly, since many works (see (Hsu and Lin, 2002) as a good example) have argued that the one-against-one method is a competitive research in multi-class SVM classification, a realization of one-against-one method is used to accomplish the task.

4.5. Simulation Experiments

4.5.1. Experiment Setup

4.5.1.1. Data Setup

In the experiments, natural images taken from real world environment are used to validate the effectiveness of proposed approach. To measure the performance of the proposed approach and better correlate with the auditory perception results obtained in Chapter 3, environmental object images of clock, phone, police car and patrol car are chosen to generate the experiment dataset. Meanwhile, the objects contained in the images are ordinary and can be frequently seen in the outdoor or indoor environment, while most of them have unique sound properties. Thus, useful visual information of objects obtained in this experiment can be fused in Chapter 5 with the acoustic information to form a higher level knowledge of specific environmental event.

Moreover, the classification model of all the objects are trained on 70 images per class while the background dictionary is learned by using the images taken in real environment including both natural and indoor scenarios. Based on the proposal of this approach, the performance of salient foreground objectness detection relies on the background dictionary while the classification performance depends on the accuracy of trained classifiers. Therefore, the images of experimental environment are added to the training examples to enable an acceptable performance of the learned background dictionary. To improve the classification performance, the training images of objects consist of different shapes as well as colors within the same category.

4.5.1.2. Experimental Protocol

Followed by the proposed salient foreground environmental object detection and classification approach, the experiments are conducted in a popular used dataset as well as real image obtained from internet. The protocol of simulation experiments of this chapter can be summarized as the following procedures:

1) Learn the overcomplete dictionary of background

There are 150 pictures which randomly chosen from internet with different colors and shapes for training the dictionary. The pictures rarely have foreground objects and are photographed from ordinary environments which can be commonly seen in human world. The learning process is conducted on the laptop with Intel i7-3630QM cores of 2.4 GHz and 8 GB internal storage, 100 iterations are deployed as a compromise of time and computational cost.

2) Train the BoW model of expected objects.

The experimental images from each category consists of 70 pictures or so, each of which is obtained randomly from internet or photographed by using iPhone 4s with a 8 megapixels camera. To provide better performance of training, all the raw pictures are clipped to ensure the expected object can be filled with the entire image. However, the background of the training pictures can be varying while the color and shape could be also different.

3) Salient environmental object detection by using visual saliency.

For a test image $I(z)$, the visual saliency detection result can be obtained by using the mentioned saliency filter approach. The detected objects will be represented by the brighten regions in the output images and the brightness indicates the degree of visual saliency. The most salient object or area will be located in the brightest region of the visual saliency map.

4) Foreground environmental object detection by using objectness measurement.

For a test image *I*(*z*), it will be reshaped into 640×640 pixel and divided into 3 scales as mentioned in section 4.4.1. For different scales, the objectness is performed within a local image patch that cut out by a sliding detection window of 8×8 pixel. Firstly, the 2-D gabor feature of this local image patch is calculated. Second, the sparse coefficient α of the calculated gabor feature vector is obtained by using sparse decomposition over the previous learned background dictionary *D*. Third, reconstruct the feature vector by applying the calculated sparse coefficient α on the dictionary *D*. Finally, the reconstruction error is calculated by using the Euclidean distance between the original feature vector and the reconstructed feature vector. Thus, the set $\varepsilon_i^{s_j}$ of reconstruction error for i^{th} patch in s_j scale is obtained. By comparing with the error threshold of ρ_{s_j}, the objectness of each patch is obtained and the patches with larger error than the threshold are considered to be potential candidates of foreground object.

5) Foreground object classification by using BoW model.

For the candidates of foreground object obtained in step 4, the DSIFT feature of each patch is calculated. By using the trained BoW model mentioned in step 2, the exact local patches in which the expected object could locate are determined in all the scales. Particularly, to eliminate the reduplicate local image patch in each scale, the exclusion ratios of non salient area in all the three scales are set to be 95%, 99.5% and 99,9% in corresponding with the scales of 16×16 pixel, 32×32 pixel and 64×64 pixel.

6) Perform the classification process in a human like way.

To provide the machine with a human-like visual object perception approach, the classification is performed by using the fusion discrimination of visual saliency and objectness detection. To be specific, the input image will be processed by the visual saliency detection first, the result of which will be applied for object classification. If the classification process is failed as no expected object is classified in the visually salient area, the objectness detection approach will be carried out.

4.5.2. Experiments Result and Discussion

In order to compare the performances of both visual saliency based method and objectness based method, the detection results derived from the popular PASCAL VOC2007 dataset (Everingham et al., 2007) are shown in Figure 18 to demonstrate the importance of using objectness property in foreground object detection.

In Figure 18, four examples from both indoor and outdoor environments have been given to demonstrate the differential results of foreground object detection by using the mentioned visual saliency detection approach and the proposed objectness detection approach. To be specific, the original images, saliency maps and objectness detection results are shown in the first, second and third row, respectively. It can be seen from the original images that, the foreground object that salient in informatics with respect to human perception characteristic in each test image can be illustrated as: two sheep in column (a), airplane and people in column (b), chairs and small sofas in column (c) and computer with keyboard in column (d).

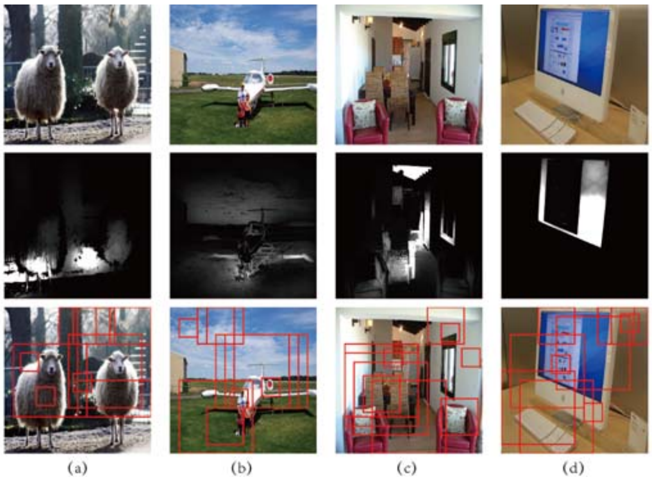


Figure 19: The foreground objects detection results of four test images.

The visual saliency detection results in the second row of Figure 18 have shown that, the salient regions in example (a) represent the grass with green color behind the sheep and a small part (i.e. legs) of one sheep (i.e. left) while the majority of the two sheep have not been detected as salient objects; the salient blob in example (b) can be hardly recognized as an airplane since the main structure of the airplane is not visually salient and the people are not detected as well; the most salient objects detected in example (c) are the door and ceiling with dark color which are less interesting as they can be considered as backgrounds, another salient region represents the table which masked by the chairs and all the chairs have not been correctly detected. The middle image from column (d) shows that the blue part of computer screen has been detected as salient region, while the entire computer and the keyboard are the expected salient foreground objects. Therefore, the result images from the second row have shown that the visual saliency detection could not extract the expected foreground objects when the objects are not visually salient but salient in informatics.

The objectness detection results of four test images are shown in the third row of Figure 18, in which the red windows with different size are detection windows used in different scales. It can be clearly seen from the result images that, despite there are a few mismatched windows that located in the background, such as the sky in column (b), the wall in column (c), the majority of all the red windows can correctly include the expected foreground objects. Since the objects within the detection windows will be considered as the candidates of foreground object, windows which only cover a small part of the object will not affect the classification process as long as the objects are covered by large windows.

Furthermore, in order to obtain both visual and acoustic information that relate to the same object or the same event for higher level heterogeneous information fusion, images taken in the real world with pre-assigned environmental objects are applied to validate the detection and classification performance of the proposed approach. The experimental results are graphically shown in Figure 19.

In this experiment, clock and phone are the expected foreground objects that being considered in the indoor experiment. Meanwhile, the police car and patrol car with different visual representations are considered to be salient in informatics and assigned to be the expected objects in outdoor environment. To ensure the test images are different from the images obtained from well-established dataset while the quality and resolution of the test image can represent the actual requirement of real world, the images of clock and phone are taken in a typical office room, while the images of police and patrol cars are randomly selected from the Internet via *Google.fr*. Notably, other objects are simultaneously appeared in the pictures which could be treated as interferences, while some of which are also visual salient to human visual perception.

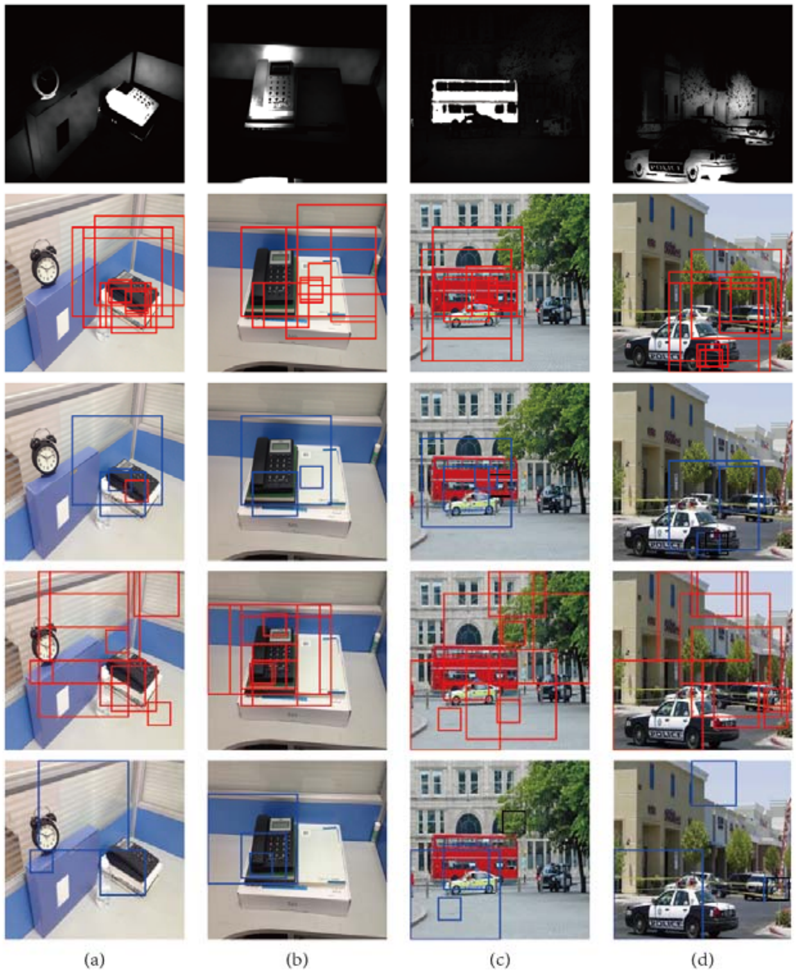


Figure 20: The salient foreground objects detection and classification results of (a) clock, (b) phone, (c) patrol car and (d) police car.

In Figure 19, the visual saliency images are demonstrated in the first row, the object detection and classification results by using visual saliency based method are shown in the second and third row, while the results of proposed objectness based approach are given in the last two rows in which the forth shows the images of objectness detection and images from the fifth row show the classification results.

The images from examples (a) and (c) of Figure 19 show that, visually salient objects are detected while foreground objects that salient in informatics can not be located, such as the clock in example (a) and the car under the tree in example (c). Though this could has little influence to the classification while the salient foreground object is not the expected object, such as the car under the tree in example (c), it still could lead to a failure in classification as shown in the third image of example (a). The forth images from both example (a) and (c) show that, all the salient foreground objects have been detected and classified. Particularly, both of the expected objects of clock and phone are classified by using objectness based approach as shown in the last image of Figure 19(a), and the detection windows in the last image of Figure 19(c) are more close to the expected patrol car compare to the visual saliency detection results in the third image of Figure 19(c). These two examples show that the proposed objectness based approach is able to successfully detect and classify the foreground objects which are salient in informatics when the expected objects are not visually salient.

Meanwhile, a test image consists of a visually salient object of phone is given in Figure 19(b). Though the third image of example (b) shows that the classification based on visual saliency approach is correct with only one small mismatched window exists, better classification result with no mismatched detection window exist could be obtained by using the proposed objectness based approach as shown in the last image of example (b). Furthermore, two cars are considered to be the foreground objects while the larger one is the expected police car as shown in the test image of example (d). It can be seen from the third and forth images of Figure 19(d) that, though the police car is detected along with another foreground car, a successful classification result can only achieved within the detection window that only cover the salient part of the police car. However, majority of the visual representation of police car can be detected by the detection window as shown in the last image of Figure 19(d), which demonstrates the effectiveness of the proposed approach. Nevertheless, there are still some mismatched detection windows exist in the results obtained by using proposed approach, the explanation for this limitation is that only a small number (N=150) of background images are applied in my work to train the background dictionary. Thus, the dictionary is not well constructed based on mass data and not all the backgrounds can be comprehensive represented by the learned dictionary. Moreover, the boundary between background and foreground is ambiguous and even subjectively different according to the differentials in visual perception system of different people.

4.6. Conclusion

In this chapter, a visual information perception approach based on the detection and classification of salient foreground environmental object is researched. Motivated by the visual perception property of human beings and the generic characteristic of object, the environmental objects are characterized as salient in informatics. In order to capture the interesting foreground objects for perception, a novel foreground object detection approach is proposed based on the objectness characteristics of objects. To achieve the detection goal of foreground object, a sparse representation based method is initially presented to obtain the objectness feature of object different from other methods. To be specific, the objectness of salient foreground object is obtained by calculating the dissimilarity between the object feature and the background dictionary based on the reconstruction error. Meanwhile, the visual saliency detection approach is considered to accelerate the detection process as the simulation of bottom-up perception mechanism when the foreground object is either salient in informatics or visually salient. Moreover, as the foreground objects exist in environment are not always visually salient to human awareness, the active searching of expected object should be also taken into account. Therefore, a fusion discrimination scheme based on the combination of visual saliency and objectness property is presented to accomplish the human-like perception task. Furthermore, the classification of salient foreground object is carried out by using the popular model of bag of visual words and the classifier of support vector machine.

Experiment results derived from the popular VOC2007 dataset show that, the proposed objectness based approach can correctly detect the foreground objects which are salient in informatics when the visual saliency detection approach failed, which demonstrates the effectiveness of proposed approach. The classification results of real world images show that, the performance of objectness based approach is competitive in detecting salient foreground object. Though the classification accuracy of proposed approach has small limitation as mismatched detection window exists in background, more accurate results are considered to be possible when comprehensive dictionary learning process is applied.

In general, the visual information awareness characteristic of salient foreground environmental object for machine can be obtained by applying the proposed approach in this chapter, while the visual perception information could be achieved to form the knowledge of environmental object's visual representation for further heterogeneous information fusion process.

Chapter 5. Heterogeneous Information Fusion Framework for Human-Like Perception of Complex Environment

5.1. Introduction

In the exploration of surrounding environment, both image and sound signals are valuable for mankind to comprehensively understand the environment. Human beings can combine the heterogeneous information obtained from sound and image channels in an efficient and intelligent way, while semantic interpretation and active perception could be achieved as well. Due to the generic differentials between image and sound signals, it is hardly possible for machines to merge these two kinds of heterogeneous information at signal level effectively. Therefore, as illustrated in Chapter 2, most the previous works were conducted based on the fusion in feature and knowledge levels.

Considering the heterogeneous information fusion in feature level, audio feature are more frequently used as the auditory cue to discriminate the visual detection in certain applications, such as abrupt change detection (Chen et al., 2014), violent scene detection (Giannakopoulos et al., 2010), speaker tracking (Li et al., 2012) and human aggression detection (Lefter et al., 2013). Particularly, (Liu et al., 2014b) proposed a fusion approach by using audio vocabulary and visual features for pornographic video detection, in which the audio is represented by bag of words (BoW) model and the energy envelope unit (EEU) based semantic representation. It should be noticed that, though results of previous works are promising in specific application scenarios, the fusion schemes proposed in these works are mainly based on the linear discrimination or conceptual decision in order to determine the final fusion result. Hence, the audio information is only considered as a complementary part of visual information while not as the representation of environment. Therefore, the audio-visual fusion technique is not suitable for accomplishing the perception task in the complex environment as the heterogeneous information of sound and image are equally important.

To overcome the limitation of applying feature level fusion, the knowledge level fusion approaches use the information derived from image and auditory channels to form a higher level knowledge of environment. However, as knowledge is considered to be higher level information about object or event, fusion methods are frequently based on the pre-designed rules of determination and inference to form the decision results by combining heterogeneous information together, such as the D-S evidence theory, fuzzy logic, fuzzy set and rough set, to name a few. As most knowledge fusion techniques can be seen as the process of decision making, they will be more suitable in dealing with diagnosis and judgment issues under specific circumstances than real environment due to the semantic complexity of perception issue. In the exploration of surrounding environment, all the perceptually interesting and valuable information are generated by various kinds of environmental objects with different visual and auditory characteristics. Consequently, novel approach should be researched in order to better interpret the event occurred in the environment.

Motivated by the mentioned shortcomings of current approaches and the practical requirement of intelligent environment perception, a novel heterogeneous information fusion framework is proposed in this chapter. The proposed framework is based on the semantic representation models of image and sound information, in which the actual and potential information are built by using the probability topic model. Meanwhile, a scene information probability model is proposed to increase the invariance ability of proposed framework in complex environment. Furthermore, the property of negative probability is taken into account in the models to better describe the environment and the heterogeneous information as well.

5.2. Proposal of Framework

The fundamental assumption of the proposed framework in this chapter is that all the environmental incidents as well as the environment itself are composed of various kinds of objects. Despite that the sound and image information are heterogeneous, the contribution of heterogeneous information derived from visual and auditory channels are equally important in semantics from the perspective of perception.

Motivated by the probability topic model (detail described in section 5.3.1) which proved to be useful in revealing the semantic composition of document and language, an adapted version of information probability model is proposed to describe the visual and acoustic objects. Thereafter, a scene information probability model is presented to demonstrate the relationships between scenes and objects. Inspired by the perceptual process of human perception ability, the proposed framework is performed by using an information probability model in which heterogeneous information are considered as components of objects with probabilities. Regarding the potential events that could be abnormally occurred in the environment, the negative property of object is initially proposed as novel extension of the probability component to provide comprehensive description of object. The diagram of proposed heterogeneous information fusion framework is presented in Figure 20 to demonstrate the process procedure.

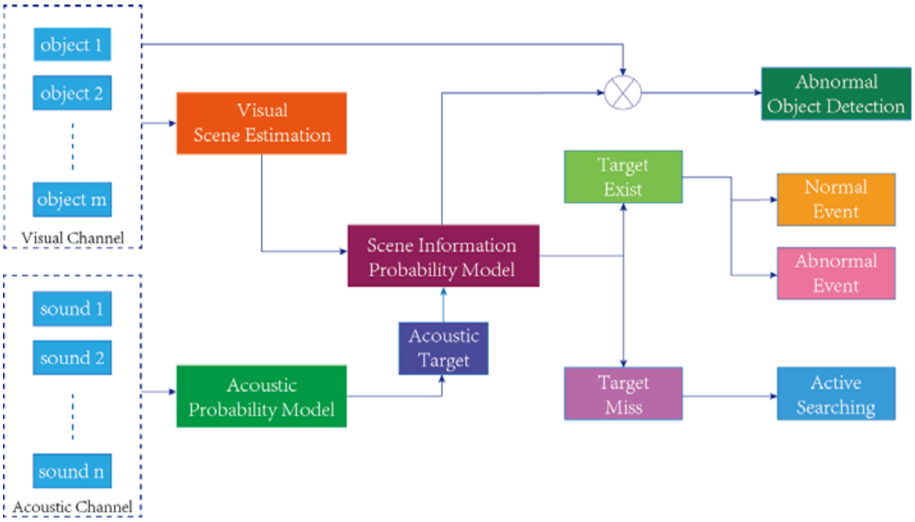


Figure 21: The general structure of proposed heterogeneous information fusion framework for environment perception.

Due to the distinctive characteristic of visual stimuli in perception, the detected objects from visual channel will be used to analyze and estimate the specific category of current environment in a scene understanding manner. Meanwhile, since only the salient environmental sound is considered to be valuable as it represents certain events occurred in the environment, the sound information will be processed by the acoustic probability model to form the knowledge of acoustic object in parallel. Thereafter, the scene information probability model is proposed as a representation of environment to uniformly combine heterogeneous information of sound and image together. By using the proposed model, the environment can be described by heterogeneous information in a joint way that can provide comprehensive explanation of either the environment or the event occurred in the environment. Therefore, the analysis and understanding of environment itself as well as the environmental event could be obtained to provide the final perception results for artificial perception of machines.

5.3. Probability Topic Model based Heterogeneous Information Representation

5.3.1. Probability Topic Model

The probability topic model (PTM) is also known as the probabilistic topic model, which was developed based upon the general idea that documents are mixtures of topics, where a topic is a probability distribution over words. It was frequently applied in the research filed of latent semantic analysis (LSA) to overcome the shortcomings of conventional Information Retrieval (IR) techniques (Wang et al., 2010a). The general purpose of using PTM is to analyze the latent content of documents and calculate the similarity between documents in an efficient way to model the document and extract the actual topics. Since variety of PTMs have been proposed as extensions to improve the traditional LSA approach (see (Blei et al., 2003), (Griffiths and Steyvers, 2003), (Griffiths and Steyvers, 2004) and (Griffiths and Steyvers, 2005) as examples of previous works), the simplest model of Latent Dirichlet Allocation (LDA) is briefly introduced as the motivation of my work.

The LDA model is proposed by (Blei et al., 2003) based on the probabilistic topic model introduced in (Hofmann, 2001). Similar to other PTMs, the intuition of LDA is that documents consist of multiple topics, which can be seen as distributions over different vocabularies. Different from other approaches, the generative process of new document in LDA model is introduced by applying a Dirichlet prior of mixture weight. Since the Dirichlet distribution is a conjugate prior for the multinomial, it can be used as prior to simplify the problem of statistical inference. In (Steyvers and Griffiths, 2007), regarding a multinomial distribution $p=(p_1,\ldots,p_T)$, the probability density of a T dimensional Dirichlet distribution can be defined by

$$Dir(\alpha_1,\ldots,\alpha_T)=\frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)}\prod_{j=1}^{T} p_j^{\alpha_j-1} \tag{64}$$

where α_j denotes the set of hyperparameter which represents the prior observation count for the number of times topic j is sampled in a document. To illustrate the generative process of LDA, an example is given to graphically demonstrate the intuitions behind LDA in Figure 21.

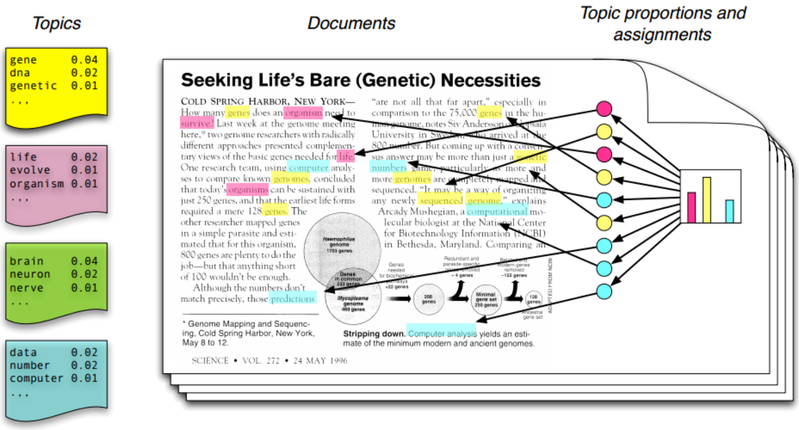


Figure 22: An example of the intuitions behind latent Dirichlet allocation (LDA). Adopted from (**Blei, 2012**).

It can be seen from Figure 21 that, the example document is a distribution over topics while each topic is a distribution over fixed vocabulary that including specific words. Therefore, the LDA model is considered to be a statistic approach which aim to capture the latent semantic meaning of documents.

5.3.2. Information Probability Model

Motivated by the general idea of LDA model, an observation can be obtained that the perception of environment shares the similar semantic structure. To be specific, when an object is regarded as a topic, the heterogeneous information can be seen as the words included in the vocabulary. Thus, the incidents that generated by the object could seen as the statistical probability problem.

Moreover, the environment can be seen as a topic as well, in which different objects with heterogeneous information compose the vocabulary for different topics. Therefore, the heterogeneous information of image and sound which correlate to one object can be treated as the probability distribution of information to form a novel representation model in knowledge level, which is named as information probability model (*IPM*) in this thesis.

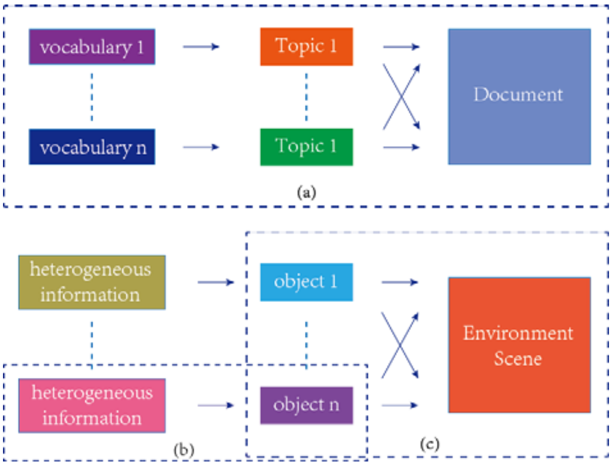


Figure 23: The similarity in semantic structure between (a) LDA model and proposed IPM model of (b) object and (c) environmental scene.

The similarity between LDA model and proposed IPM model is shown in Figure 22, in which the *IPMs* of both object and environmental scene are presented in a concatenated form.

Specifically, the general structure of simplified LDA model is shown in Figure 22(a) in which the probability relationship has been replaced by the arrows, while the proposed IPM model including objects and environmental scene is shown in the part of Figure 22(b) and (c). Note that, different colors of vocabulary and heterogeneous information denote the differential of properties, and different topics and objects are indicated by the differentials in color as well. Particularly, both visual object and acoustic object are taken into account as the sources of heterogeneous information in the IPM model in order to generate comprehensive knowledge for perception.

5.3.3. Heterogeneous Information Modeling

For an ordinary environmental object that can be found in a typical environment of human beings, the *IPM* model is based on the heterogeneous information obtained from both image and sound channels. Similar to human perception characteristic, it is widely accepted that image plays a vital role in the perception of artificial machines. By using visual information, perception work such as environment scene analysis and foreground object detection can be achieved efficiently. However, the importance of sound information should be equally emphasized as it could provide comprehensive information about occurred environment events and existed environmental objects as well. Another advantage of applying sound information is that human beings will use sound as an indicator to know the property and characteristic of object for knowledge acquisition, which make sound signal becomes the probabilistic information for object. The main purpose of heterogeneous information modeling is to regard both image and sound as probabilistic information of object, which partly describe the characteristic of object in perception. To be specific, image can be seen as a baseline representation which passively provides the visual information, while sound is considered to be more actively in providing event-related interactive information. Therefore, the *IPM* of i^{th} object can be defined as

$$IPM_i = \left\{ \sum_{j=1}^{m} V_{i,j} \mid V_{i,j}^{P} \cup \sum_{j=1}^{n} S_{i,j} \mid S_{i,j}^{P} \right\} \square\, Obj_i \tag{65}$$

where $\sum_{j=1}^{m} V_{i,j} \mid V_{i,j}^{P}$ is a set of *m* visual information $V_{i,j}$ of i^{th} object in which each $V_{i,j}$ has a corresponding probability of $V_{i,j}^{P}$ in j^{th} environment, $\sum_{j=1}^{n} S_{i,j} \mid S_{i,j}^{P}$ is a set of *n* acoustic information $S_{i,j}$ of i^{th} object in which each $S_{i,j}$ has a probability of $S_{i,j}^{P}$ in j^{th} environment. Similar to the aforementioned PTM, the result of IPM_i is defined in the form of perceptual knowledge Obj_i which indicates the name of object. The *IPM* of object in Eq. (65) shows that each object existed in the environment can be modeled as the combination of heterogeneous probability information.

Meanwhile, since the same information obtained from either visual or acoustic channel could be generated by multiple similar objects, it is therefore important to model the heterogeneous information with respect to different objects as

$$V_j = \prod_{i=1}^{N} IPM_i \mid P_{V_j}^{i} \tag{66}$$

$$S_j = \prod_{i=1}^{N} IPM_i \mid P_{S_j}^{i} \tag{67}$$

in which V_j and S_j represent the sets of appearance probability of j^{th} visual and acoustic information in all the *N* IPMs with probabilities of $P_{V_j}^{i}$ and $P_{S_j}^{i}$ in i^{th} IPM, respectively.

Consequently, by calculating the probability of appearance of j^{th} heterogeneous information according to actual condition and situation of environment, the generator candidates of perceived heterogeneous information can be obtained. Notably, $P_{V_j}^{i}$

and $P_{S_j}^i$ can be either scene-independent or scene-related based on the practical requirement of different perception tasks. For example, Eq. (66) and Eq. (67) can be seen as the simplest case when *i*=1, which means that heterogeneous information are uniquely appeared in one object and each object generate its own set of heterogeneous information. The general assumption of applying probability property to describe the appearance of heterogeneous information is that, the same heterogeneous information can be generated by certain kinds of environmental objects more frequently than ever before, especially when machines with screens have been invented such as phone, TV and computer. Consequently, despite that the perceived information could indicate the traditional object with high probability, there do exist the possibility that other objects could generate the same information as well which could leads to different perceptual knowledge of environment.

5.3.4. Scene Information Probability Model

The environment is vivid and realistic to our minds because variety of events is occurring continuously. From perception point of view, the environment will be more valuable and interesting for exploration if the scene is not familiar or surprised events exist. This perceptual motivity can be characterized as the curiosity of human beings. Therefore, the analysis and understanding of the environment scene is important in the perception of events and environment itself.

To better fuse the heterogeneous information with respect to different objects and environments in a more coordinated way, the scene information probability model is proposed to combine information, objects and environment scene together based on the proposed IPM model. Considering a typical environment consists of *N* objects, the scene information probability model (sIPM) can be defined as

$$sIPM_i = \prod_{k=1}^{N} Obj_k \mid P_{obj}^{i,k} \qquad \textbf{(68)}$$

where Obj_k denotes the k^{th} object that could appeared in the environment, $P_{obj}^{i,k}$ is the appearance probability of k^{th} object in i^{th} scene. Note that, Eq. (68) shows that scene information is the combination of probability distribution of *k* objects that Obj_k can be represented by using the IPM defined in Eq. (65) which reforms Eq. (68) as

$$sIPM_i = \prod_{k=1}^{N} \left\{ \left[\sum_{j=1}^{m} V_{k,j} \mid V_{k,j}^{P} \cup \sum_{j=1}^{n} S_{k,j} \mid S_{k,j}^{P} \right] \mid P_{obj}^{i,k} \right\} \quad \textbf{(69)}$$

It can be derived from Eq. 5-6 that, the sIPM of i^{th} scene will be transformed into simple perception models of using homogeneous information. For example, sIPM will become the model applied in the traditional scene analysis research when acoustic probability information is absence, otherwise becomes the auditory scene analysis model when acoustic information is missed. Therefore, sIPM model can be seen as a generalized model of environmental information with probability distribution, which is able to combine the information obtained from heterogeneous channels, objects and scenes together to form a more comprehensive model of environment.

5.4. Heterogeneous Information Fusion based Environment Perception

5.4.1. Motivation

For environmental objects existed in real environment, there could be a great possibility that structural saliency problem is evitable in the perceptual relationship between foreground object and background scene. For example, the computer can be seen as salient object in a bathroom while a blower is salient in an office environment. This saliency property can not be generated by the traditional contrast operation in either image or sound channel but by a structural saliency characteristic defined by the conflict of perceptual knowledge between objects and scenes. The notion of structural saliency used in this chapter is different from the term used in the field of classic visual saliency research, which denotes the global saliency property in general. Inspired by the human awareness ability, structural saliency property characterizes the dissimilarity of perceptual knowledge of surroundings in a higher level corresponds to different environments as well as the objects, which could be seen as the fundamental motivation of curiosity of us during the perception process.

Motivated by the mentioned idea, environmental objects and events exist in the environment can be categorized into two kinds which are normal and abnormal. The term of normal means that the detected objects are expected to appear in current scene and the occurred events can be interpreted by the sIPM of scene, while the term of abnormal describes the rarity of appearance probabilities of objects in current scene as well as the events. It should be noticed that, as objects and events of environment can be considered as the probability distribution of heterogeneous information in proposed IPM and sIPM, the research issue of environment perception can be seen as the fusion of probability distribution of information by taking both normal and abnormal objects and events into account simultaneously.

Therefore, the proposed sIPM model should be modified in order to improve the perception accuracy and robustness when abnormal objects and events exist.

5.4.2. Negative Property based Complex Scene Modeling

The basical assumption of negative property modeling could be easily concluded from the proposed sIPM. To be specific, environment scenes will consist of certain set of visual and acoustic objects with limited numbers respectively that objects could be rarely seen in one environment scene while frequently appeared in other scenes. Thus, the appearance rarity of object in different environment scenes is defined as negative property of probability information.

Similar to the definition of IPM and sIPM, the negative probability information of L structural salient objects in the i^{th} envrionment scene can be given based on Eq.

(68) as

$$nsIPM_i = -\prod_{l=1}^{L} Obj_l \mid P_{nobj}^{i,l} \quad \textbf{(70)}$$

where Obj_l is the IPM of l^{th} structural salient object with the probability of $P_{nobj}^{i,l}$ in i^{th} environment scene. However, as the value of probability can not be negative, the minus sign in Eq. (70) is merely a representation symbol that indicates the negative characteristic of information. The perceptual concept of Eq. (70) is that the rarities of structural salient objects in scenes are described by the value of not-appear probability while the negative symbol is used as the representation constraint. By using negative probability information, the expected objects and structural salient objects of a scene can be distinguished which makes the detection of abnormal object possible. This process can be visualized as a separation operation which shown in Figure 23.

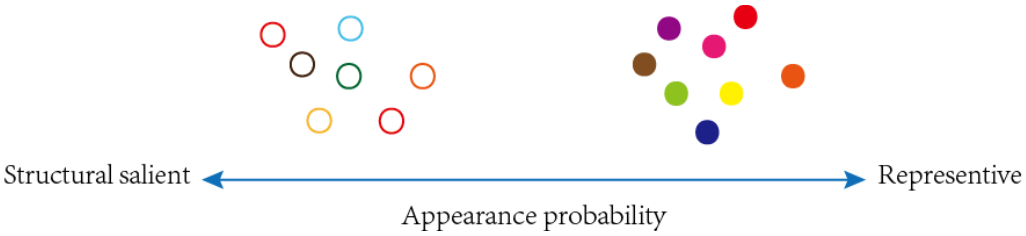


Figure 24: The appearance probability of representive objects (solid) and structural salient objects (hollow) in one scene.

As shown In Figure 23, the horizontal axis represent the appearance probability of objects in a scene where representive objects of the scene are in the right side and structrual salient objects are in the left. For different scenes, the distribution in Figure 23 will be various as both representive objects and structural salient objects could be different. Hence, complex scenes could be practically modeled by using the proposed nsIPM and previous mentioned sIPM as complementary descriptions, then the unified scene information probability model for i^{th} scene can be given as

$$usIPM_i = sIPM_i \cup nsIPM_i \quad \textbf{(71)}$$

where the operator "$\bigcup$" denotes the assemble operation. It should be emphasized that the unified scene information probability model $usIPM_i$ of i^{th} environment scene can not be seen as the calculation result but actually is a well defined model based on the human perception knowledge.

When the scene of environment is determined by using sIPM, the corresponding usIPM will be chosen as the database for perceptual analysis. This is quite similar to the perception process of human beings, in which the scene is mainly analyzed and determined based on the visual perception information. In the analysis of a visual scene, the probability information of objects' appearances are used while the joint probability distribution of objects will determine the final result.

Once the environment scene is determined, a detailed corresponding model of the environment will be generated based on the perceptual knowledge obtained from previous experiences of learning. Accordingly, the components of proposed uIPM could be given by human experience or experimental knowledge and perhaps by learning in further steps.

5.4.3. Heterogeneous Model Fusion based Perception

Considering the ordinary perception requirement in most of the environment for machines, heterogeneous information of sound and image should be merged within a fusion framework to form higher level knowledge about environment scenes as well as the occurred events. Based on the previous proposed probability information model of objects and scenes, the perception of environment could be achieved by combining the negative probability information of scenes in a fusion manner as shown in Figure 20.

Assume that there are N_v visual objects and N_a acoustic objects exist in one environment scene. Motivated by the formulations presented in Eq. (65) to Eq. (71), the mathematical interpretation of the fusion process can be given in the following steps:

Step 1: IPM determination for the j^{th} visual object and acoustic object, then target object is obtained.

$$V_j = \prod_{i=1}^{N_v} IPM_i \mid P_{V_j}^i = Obj_j^V \text{ subject to } \arg\max_{i=1,\ldots,N_v} P_{V_j}^i \tag{72}$$

$$S_j = \prod_{i=1}^{N_a} IPM_i \mid P_{S_j}^i = Obj_j^S \text{ subject to } \arg\max_{i=1,\ldots,N_a} P_{S_j}^i \tag{73}$$

Step 2: Scene extrapolation based on sIPM by using the maximum probability distribution of V_j combination, then target scene $sIPM_i$ is obtained.

$$\arg\max_{j=1,\ldots,N_v} \prod V_j \text{ subject to } sIPM_i = \prod_{j=1}^{N_v} \left\{ \left[Obj_j^V \mid V_j \right] \mid P_{obj}^{i,j} \right\} \tag{74}$$

Step 3: Assign usIPM according to sIPM$_i$, and find the abnormal visual object by using a probability threshold τ_p based on the corresponding nsIPM$_i$.

$$\text{Find } \left| P_{nobj}^{i,j} \right| > \tau \text{ for } \forall Obj_j^V \in nsIPM_i \tag{75}$$

Step 4: Determine normal environmental event.

$$\text{Find } Obj_j^V \text{ in } sIPM_i \text{ subject to } \forall Obj_j^S \in sIPM_i \tag{76}$$

Step 5: Abnormal event perception.

$$\text{Find } Obj_j^V \text{ in } sIPM_i \text{ subject to } \forall Obj_j^S \in nsIPM_i \tag{77}$$

Step 6: Further processing for unknown acoustic objects.

If $\forall Obj_j^S \notin usIPM_i$, perform an active searching by moving camera or learning the new characteristic of the object such as visual representation.

Consequently, the major task of environment perception can be achieved by the mentioned steps. However, these steps can be performed in a more sophisticated way by considering multiple environment scenes.

5.5. Experimental Validation

5.5.1. Experiment Setup

To validate the effectiveness of proposed probability information models as well as the heterogeneous information fusion framework, typical perception tasks are built in order to cover the characteristics of classical environment scenes. Three different scenes are considered in the experiment, which are home, office and outdoor. For each environment scene, the mentioned uIPM is built based on the experience knowledge.

Based on the perceptual knowledge obtained from ordinary perception scenarios, the environment scene can be divided into four general cases as shown in Table 4 regard the appearance of heterogeneous information of object.

		Visual Information	
		Yes	**No**
Acoustic Information	**Yes**	Case 1	Case 2
	No	Case 3	Case 4

Table 4: The categorization of cases for different perception tasks.

As demonstrated in Table 4, Case 1 is the typical case represents the common situation of environment events, Case 2 can be seen as classic auditory scene analysis while visual information is absence, Case 3 can be treated as traditional scene analysis since no salient sound information corresponding to visual object has been obtained, and Case 4 will be considered pointless as neither visual nor acoustic information exists. Therefore, both Case 2 and Case 3 could be seen as the individual perception tasks in single information channel similar to the traditional research works.

Moreover, the representive objects and the objects with negative property for each environment scene are given in Table 4 for each scene.

usIPM	**Representive Objects (P_{obj})**	**Negative Objects (P_{nobj})**
Home	bed(0.95), phone(0.85), clock(0.9), dog(0.5), TV(0.85), sofa(0.8), table(0.8),…	police car(-0.9), fire(-0.9), alarm(-0.85), copier(-0.2), umbrella(-0.8), motorcycle(-0.8),…
Office	phone(0.85), cellphone(0.8), copier (0.9), computer(0.9), sofa(0.4), clock(0.7),...	police car(-0.9), fire(-0.9), alarm(-0.85), dog(-0.75), bed(-0.9), motorcycle(-0.8),…
Outdoor	dog(0.85), police car(0.75), ambulance(0.85), table(0.5), umbrella(0.8), motorcycle(0.8),…	clock(-0.8), bed(-0.9), computer(-0.8), TV(-0.7), copier(-0.8), sofa(-0.9)

Table 5: The composition of typical representive and negative objects regarding different environmental scenes.

In Table 5, several typical environmental scenes are given as examples including representive objects and negative objects. The values in parentheses of objects denote the probability of appearance while the minus sign represents the negative property of object instead of a negative value of probability. Meanwhile, the ellipsis symbols in Table 5 represent that more objects can be considered as probable candidates exist in the scene and the presented objects are limited examples in the scene.

For the purpose of comprehensively validate the effectiveness of proposed fusion framework, simulated perception task correlates to each case in one of the presented three environment scenes will be designed while ignoring Case 4 due to mentioned reason.

5.5.2. Experiments Result and Discussion

Experiment task 1: Case 1 in the home environment.

Assume the visual information of bed, clock, sofa, table, umbrella and TV have been detected while the acoustic information of clock is also perceived at the same time. Based on the IPM model, the object information could be obtained as the visual information will achieve the highest probability in the corresponding IPM of object as

$$V_j = IPM_j \square Obj_j^V \quad \text{or} \quad Obj_j^V \mid V_j = 1$$

Then the environment scene can be determined by using $sIPM$. For the three scenes, the probability information for each scene is calculated based on Step 2 as

$$sIPM_i = \prod_{j=1}^{6} \left\{ \left[Obj_j^V \mid V_j \right] \mid P_{obj}^{i,j} \right\} \quad i = 1,2,3$$

The determined scene will be assigned to $sIPM_i$ model with largest information probability value. The results are $sIPM_1$=3.5, $sIPM_2$=0.5 and $sIPM_3$=-1.9. It is clearly to see that the scene in this task should be characterized as home environment. Then the corresponding *usIPM* of home environment listed in Table 4 will be provided.

Thereafter, for the acoustic information of clock, the acoustic IPM could be also determined as

$$S_{clock} = \left\{ IPM_{clock} \mid P_{clock}^{clock}, IPM_{phone} \mid P_{clock}^{phone}, IPM_{TV} \mid P_{clock}^{TV} \right\}$$

$$= \{ Obj_{clock} P_{clock}^{clock}, Obj_{phone} P_{clock}^{phone}, Obj_{TV} P_{clock}^{TV} \}$$

By finding the largest P value within the set of S_{clock}, the acoustic object which generates the sound of clock can be determined. It is very important to emphasize that the location information of the sound is acquiesce as the sound source localization can be performed efficiently based on current approach, thus the sound information can be correlated to a specific visual object based on the probability P_{clock} for all the potential object candidates. When the acoustic object is determined, for example, the sound of clock has been assigned to the object of clock, environmental event of clock is ringing within current field of vision (FOV) can be perceived based on the process in Step 3.

Furthermore, the abnormal environmental event detection can be achieved based on the process mentioned in Step 5 as well, when the source object of salient acoustic information has been determined with respect to different scenes and IPMs in various conditions. However, the inference and calculation process are similar to the designed perception task.

Experiment task 2: Case 2 in the office environment.

Considering the second case listed in Table 3, the general setup can be featured as one acoustic target is perceived but the visual representation can not be found. There are two perceptual knowledge accounts this case. The first explanation is that acoustic target is not in the current FOV with a high probability that active searching in other FOVs should be put out; the second perceptual knowledge is that this acoustic target could be the representation of an abnormal environmental event.

1) Assume the detected acoustic information is the ringing sound of cellphone that does not appear in the current FOV of office scene. Since the visual information has been perceived, the determination of *sIPM* and *usIPM* could be easily obtained by using the similar procedure presented in experiment task 1. Therefore, the *sIPM* and *usIPM* of office environment are assumed to be provided.

For the detected acoustic information S_{clock}, the *IPM* of potential object is given as

$$S_{clock} = IPM_i \mid P^i_{S_{clock}} = Obj^S_{clock}$$

The above equation denotes the probability information of acoustic object clock in all the candidates of representive objects. By finding the maximum probability, the acoustic object can be determined. To simplify the calculation, the candidate of clock sound is defined as maximum when visual object is clock Obj^S_{clock}. Thereafter, the knowledge of scene can be achieved by applying the proposed *usIPM* in two stages. Based on the process in Step 3, the objective of first stage is to detect whether the acoustic object of clock has a visual representation corresponds to existed object that belongs to *nsIPM* as

$$\text{Find } \left|P^{office,clock}_{nobj}\right| > \tau \quad \text{for} \quad \forall Obj^V_{clock} \in nsIPM_{office}$$

Here, as the object clock does not belong to *nsIPM* of office scene and in fact is the representive object of the scene, the perceptual knowledge can be generated that this environmental event occurs beyond the current FOV. Consequently, the direction of moving current FOV of camera based on other information, such as the location of sound, should be given in the second stage in order to find the expected object.

2) Assume the salient acoustic information is the alarm sound. Then the *IPM* of object candidate can be generally given as

$$S_{alarm} = \prod IPM_i \mid P^i_{S_{alarm}} = Obj^S_{alarm}$$

However, due to the distinctive representation ability of acoustic information for environmental event, it is considered to be more effective in perceiving the abnormal event by firstly comparing the acoustic information with negative property defined in *nsIPM*. Therefore, perceptual knowledge of this situation can be obtained based on the assigned nsIPM of office scene as if

$$Obj^S_{clock} \in nsIPM_{office}$$

Notably, if the acoustic information is the probability information of an object's *IPM*, the probability information should be taken into account to yield the perceptual knowledge with respect to the corresponding *sIPM*.

Experiment task 3: Case 3 in the outdoor environment.

As illustrated in Table 3, Case 3 represents the perception situation that no salient acoustic information exists. Thus, the perception task becomes the research issue of traditional visual scene analysis. Nonetheless, the perception task of abnormal visual object detection can be also performed by using the proposed framework.

Considering the outdoor environment with j^{th} representive objects, assume the abnormal visual object is bed. The detection of abnormal visual object can be carried out based on the Step 3 of proposed framework as if

$$P^{outdoor,bed}_{nobj} < 0$$

However, more complex situations should be noticed when machines with screen exist in the environment, mismatched result could be obtained as they might generate the same visual representation but do not correspond to the real object. In real world application, it should be combined with other information to achieve more accurate result.

5.6. Conclusion

In this chapter, the knowledge level fusion framework based on heterogeneous information of sound and image is initially proposed to achieve complex perception tasks. To be specific, inspired by the probability topic model which comprehensively reveals the semantic relationships among document, topics and words, an information probability model (IPM) is originally proposed to better describe the probabilistic characteristic of heterogeneous information obtained from environment objects.

Thereafter, the scene information probability model (sIPM) of environment is presented to combine heterogeneous information for perception. By regarding both visual and acoustic representations of objects as heterogeneous information with probabilistic characteristic, the existences of objects can be treated as the calculation and inference of maximum information probability distribution with respect to scenes. Meanwhile, as objects can be subjectively characterized into normal and abnormal according to different environments, the abnormal objects as well as events can be effectively perceived in an efficient way. Motivated by the saliency mechanism, the difference between normal and abnormal is defined as the structural saliency property between objects and scenes. In order to represent this important perceptual feature and form higher level knowledge in the perception of environment, the negative property of scene is proposed to illustrate the complexity in composition of objects as well as correlated information in the environment. By combining the proposed IPMs of object and scene together, the obtained heterogeneous information is processed based on the proposed heterogeneous model fusion framework.

There are four major perception tasks could be achieved by using the proposed framework which are the perception of abnormal object, normal environmental event, abnormal environmental event and exploration of unclear object, respectively. Due to the usage of heterogeneous information probability model and negative property, various kinds of situations which could happen during perception can be correctly processed. To validate the effectiveness of proposed models as well as the fusion framework, three typical perception tasks are designed with respect to three classic environments. The experimental results of simulated perception tasks have shown that the proposed models can represent the probabilistic characteristics of heterogeneous information of objects and scenes while accomplish the perception tasks with positive perceptual knowledge.

It can be also supported by the experiments that the proposed fusion framework could cover most of the perception requirements. Particularly, multiple environmental events could be distinguished by applying heterogeneous information in an individual way or in a joint way. Therefore, the proposed approach is considered to be promising for achieving intelligent perception ability in complex environments.

General Conclusion

Conclusion

The environment perception is considered to be an important issue for intelligent machines, yet has not been well researched. Since the artificial consciousness is not practically possible to be realized on autonomous machines, an alternative choice of artificial awareness ability is taken into account and developed based on the fusion and processing of heterogeneous information in this thesis. As discussed in Chapter 1, the biological motivation of my research work comes from the saliency principle of bottom-up mechanism, which has been widely used in both human visual and acoustic perception systems. In this thesis, the saliency principle is applied not only as the general concept of approaches but also as a universal rule to concatenate different parts of my work in order to simulate the human perception process in an autonomous way.

According to the comprehensive review works of state of the art techniques that illustrated in Chapter 2, the saliency principle allows human beings to perceive the information of environment objects and events in a highly efficient way, thus makes the saliency based approach a possible solution to effectively simulate the human perception procedure and provide artificial awareness ability to machines. However, similar to the bio-inspired process of perception, computational models that used for saliency detection in both acoustic and visual channels could face many difficulties when applied in machines. Firstly, due to the complexity of acoustic environment as well as the variety of environmental sounds, very little research in the filed of sound signal processing has been done to step forward on the topic of environmental sound perception. Secondly, as valuable and interesting environment objects are not always visually salient, it could cause a sharp decrease in perception performance when applying visual saliency detection in real world scenarios. The explanation of this limitation is that, most of the expected or interesting environmental objects should be seen as salient in semantics as they carry more information than the background for perception. Consequently, traditional visual saliency approaches could not be used individually for visual information perception in real environment. Thirdly, sound and image information are very difficult to be merged together because of the fundamental differences in their natural properties. Though audio-visual fusion approaches have shown some achievements in the feature and knowledge level fusion, heterogeneous characteristics of sound and image still could not be processed within a unified fusion framework, not to mention providing a higher level knowledge about environmental objects and events for artificial machines.

In this thesis, the heterogeneous information of environment is processed based on the saliency principle to perceive the salient acoustic objects and the visual objects that salient in informatics. Due to the unconstructed characteristic of environmental sound, acoustic saliency properties of both time and frequency domains are computed to form the final saliency detection result. Meanwhile, to improve the performance and accuracy in perceiving the most salient sound, a novel bio-inspired computational model of inhibition of return (IOR) is initially proposed. By combing the IOR model evaluation result and spatial-temporal saliency features derived from MFCC and PSD of sound, environmental sounds which salient to human awareness can be detected. Then, the acoustic knowledge of salient environmental sounds could be obtained by a two level classification approach, in which the feature of spectral energy distribution is proposed based on the fuzzy set theory to simulate the fuzzy categorization ability of humans. The accuracy of acoustic saliency detection as well as the effectiveness of environmental sounds classification by using proposed acoustic perception approach has been supported by the results of real world experiments, which has demonstrated significant improvement on other traditional methods and great potential in real world applications.

On the other hand, as visually salient objects existed in the environment could not provide valuable information for perception in certain conditions; I extend the notion of saliency from visual domain to semantics domain. Motivated by this extension, the foreground objects are considered to be more salient than background in semantics, as they can provide more interesting information that crucial to perception for machines. Thus, the accurate detection and classification of foreground environmental object can be considered as a practical way of realizing the artificial awareness ability of visual perception. To accomplish the foreground objects detection task, I proposed a novel approach that combines the objectness characteristic of foreground object and visual saliency detection together to simulate the visual perception process of human beings in Chapter 4. Different from other traditional methods, I use the sparse representation error to detect the foreground objectness property. To be specific, objectness is gained by computing the dissimilarity between object feature and the background dictionary based on the reconstruction error. Moreover, I apply the state of the art classification model for foreground object classification in which classic image features are used. The results of experiments have shown the effectiveness of proposed approach as well as the similarity between artificial awareness ability and human perception property.

Regarding the ultimate research goal of this thesis, which is the heterogeneous information fusion for environment perception of intelligent machines, I developed a novel fusion framework based on the probabilistic characteristics of heterogeneous information in Chapter 5. Inspired by the traditional probability topic model (*PTM*) which frequently used in semantic analysis of document, I regard the heterogeneous information of sound and image as the probability distribution of object's information and initially proposed an information probability model (*IPM*) to comprehensively describe the probabilistic property. Thereafter, I extend the proposed IPM of objects to *sIPM* of environment scene by regarding it as the probability distribution of objects, thus associate the heterogeneous information with the environment scene via objects for machine awareness. By introducing the concept of structural saliency, abnormal objects and events can be characterized as the negative property in the proposed *sIPM*, which leads to the proposition of negative information probability model (*nsIPM*) and unified scene information probability model (*usIPM*). Based on the proposed models, a novel fusion framework is further built for complex environment scene analysis as well as the perception of both normal and abnormal objects and events. Several cases generated from the typical scenarios of real world are applied to validate the proposed heterogeneous information fusion framework. The experiment results have shown that, multiple perception tasks can be accomplished under different circumstances based on the heterogeneous information, while the proposed *IPM*, *sIPM*, *nsIPM* and *usIPM* have been proved to be effective in processing and fusing heterogeneous information.

Perspectives

The saliency driven awareness approach based on heterogeneous information of sound and image for environment perception that has been proposed and developed in this thesis proved to be an effective and practical way of realizing artificial awareness ability on machines. Although the conducted research work has achieved competitive results for machine awareness, perspectives from several aspects should be addressed to illustrate the future work of my thesis.

Regarding the salient information extraction in acoustic domain, though multiple acoustic saliency features as well as a computational IOR model have been proposed for salient environmental sounds detection, the proposed approach still could have the possibility that it will not correctly locate the salient sounds because of the variety of environmental sounds. Meanwhile, the classification of similar environmental sounds is also considered to be a difficult research issue in acoustic perception as machines can not distinguish similar sounds with different time-frequency properties from one category. Thus, an interesting direction of future research would be the application of dynamic time-frequency structure. This novel characteristic can be seen as an analysis technique that inspired by traditional time-frequency analysis methods, in which the sound signal will be analyzed within an adaptive or fixed window frame to reveal the dynamic structure. By applying the proposed approach on each local window frame, a more comprehensive description of sound signal could be achieved to demonstrate the unconstructed characteristic of environmental sound.

As for the image information acquisition which based on the objectness measure, it has been supported by the experimental results in Chapter 4 that, it is a possible way of detecting the foreground objects that salient in semantics by calculating the sparse representation based reconstruction error. However, the performance of this approach highly relies on the generated background dictionary. In this thesis, I only use 150 pictures for background dictionary learning in which just 100 iterations are applied to decrease the computational cost. The underlying perspective of this aspect of work could be the expansion in training example and accuracy improvement in background dictionary learning.

The major research work in this thesis is the fusion of heterogeneous information for environment perception. By applying the proposed *IPM*, *sIPM*, *nsIPM* and *usIPM*, multiple perception tasks could be successfully accomplished for machine awareness. The experimental cases designed in Chapter 5 have demonstrated the effectiveness of proposed fusion framework in dealing with different real world scenarios. As inferred by the results, the performance of proposed fusion framework in dealing with natural environment should be tested in real world as well. As for the perspective of the future work, validation on real intelligent machines such as intelligent robot should be taken into account. Since the intelligent perception and cognition of robot laboratory will be established in the near future, this part of work is expected to be conducted soon.

Publications

Wang, J., Zhang, K., Madani, K., & Sabourin, C. (2013, October). A visualized acoustic saliency feature extraction method for environment sound signal processing. In *TENCON 2013-2013 IEEE Region 10 Conference (31194)* (pp. 1-4). IEEE.

Wang, J., Zhang, K., Madani, K., & Sabourin, C. (2013, November). Heterogeneous information saliency features' fusion approach for machine's environment sounds based awareness. In *Awareness Science and Technology and Ubi-Media Computing (iCAST-UMEDIA), 2013 International Joint Conference on* (pp. 197-205). IEEE.

Wang, J., Zhang, K., Madani, K., & Sabourin, C. Salient Environmental Sound Detection Framework for Machine Awareness. *Neurocomputing*. Available online 28 October 2014. DOI: http://dx.doi. org/10.1016/j.neucom.2014.09.046.

Wang, J., Zhang, K., Madani, K., & Sabourin, C. Multi-Scale Features based Salient Environmental Sound Classification for Machines Awareness. In *6th International Conference on Awareness Science and Technology (iCAST 2014), 2014*.

Bibliography

[**Abbott, et al. 2012**] Abbott, C. C., Merideth, F., Ruhl, D., Yang, Z., Clark, V. P., Calhoun, V. D., Hanlon, M. F., & Mayer, A. R. (2012). Auditory orienting and inhibition of return in schizophrenia: a functional magnetic resonance imaging study. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *37*(1), 161-168.

[**Achanta, et al. 2008**] Achanta, R., Estrada, F., Wils, P., & Süsstrunk, S. (2008). Salient region detection and segmentation. In *Computer Vision Systems* (pp. 66-75). Springer Berlin Heidelberg.

[**Achanta, et al. 2009**] Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009, June). Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 1597 -1604). IEEE.

[**Achanta, et al. 2010**] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2010). *Slic superpixels. École Polytechnique Fédéral de Lausssanne (EPFL)*. Tech. Rep, 149300.

[**Adami 2006**] Adami, C. (2006). What do robots dream of?. *Science*, *314*(5802), 1093-1094.

[**Aharon, et al. 2006**] Aharon, M., Elad, M., & Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, *54*(11), 4311-4322.

[**Ahonen, et al. 2004**] Ahonen, T., Hadid, A., & Pietikäinen, M. (2004). Face recognition with local binary patterns. In *Computer vision-eccv 2004* (pp. 469-481). Springer Berlin Heidelberg.

[**Ahonen, et al. 2006**] Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *28*(12), 2037-2041.

[**Aleksander 2001**] Aleksander, I. (2001). The Self 'out there'. *Nature*, *413*(6851), 23-23.

[**Alexe, et al. 2010**] Alexe, B., Deselaers, T., & Ferrari, V. (2010, June). What is an object?. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (pp. 73-80). IEEE.

[**Alexe, et al. 2012**] Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the objectness of image windows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(11), 2189-2202.

[**Ali Shah, et al. 2013**] Ali Shah, S. A., Bennamoun, M., Boussaid, F., & El-Sallam, A. A. (2013, February). Automatic object detection using objectness measure. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on* (pp. 1-6). IEEE.

[**Allen 1994**] Allen, J. B. (1994). How do humans process and recognize speech?. *Speech and Audio Processing, IEEE Transactions on*, *2*(4), 567- 577.

[**Anderson 2005**] Anderson, J. R. (2005). *Cognitive psychology and its implications*. Macmillan, 519.

[**Anwer, et al. 2011**] Anwer, R. M., Vázquez, D., & López, A. M. (2011). Opponent colors for human detection. In *Pattern Recognition and Image Analysis* (pp. 363-370). Springer Berlin Heidelberg.

[**Anwar, et al. 2014**] Anwar, S., Zhao, Q., Manzoor, M. F., & Ishaq Khan, S. (2014). Saliency detection using sparse and nonlinear feature representation. *The Scientific World Journal*, *2014*, 1-16.

[**Arrabales, et al. 2009**] Arrabales, R., Ledezma, A., & Sanchis, A. (2009). CERA-CRANIUM: A test bed for machine consciousness research. *International Workshop on Machine Consciousness*, Hong Kong, 1-20.

[**Attneave 1954**] Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, *61*(3), 183.

[**Baars 2002**] Baars, B. J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in cognitive sciences*, *6*(1), 47-52.

[**Ballard 1981**] Ballard, D. H. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern recognition*, *13*(2), 111-122.

[**Barlow 1961**] Barlow, H. B. (1961). Possible principles underlying the transforma -tion of sensory messages. *Sensory communication*, 217-234.

[**Barlow 1972**] Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology. *Perception*, *1*(4), 371-394.

[**Bay, et al. 2006**] Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. In *Computer Vision–ECCV 2006* (pp. 404-417). Springer Berlin Heidelberg.

[**Beltrán-Márquez, et al. 2012**] Beltrán-Márquez, J., Chávez, E., & Favela, J. (2012). Environmental sound recognition by measuring significant changes in the spectral entropy. In *Pattern Recognition* (pp. 334-343). Springer Berlin Heidelberg.

[**Beritelli, et al. 2008**] Beritelli, F., & Grasso, R. (2008, December). A pattern recognition system for environmental sound classification based on MFCCs and neural networks. In *Signal Processing and Communication Systems, 2008. ICSPCS 2008. 2nd International Conference on* (pp. 1-4). IEEE.

[**Blei, et al. 2003**] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

[**Blei 2012**] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

[**Bosch, et al. 2007**] Bosch, A., Zisserman, A., & Munoz, X. (2007, October). Image classification using random forests and ferns. In *Computer Vision, IEEE International Conference on* (pp. 1-8). IEEE.

[**Brandt 2010**] Brandt, J. (2010, June). Transform coding for fast approximate nearest neighbor search in high dimensions. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (pp. 1815-1822). IEEE.

[**Broadbent 1958**] Broadbent, D. E. (1958). The selective nature of learning. In *Perception and communication*. Elmsford, NY, US: Pergamon Press, 244-267.

[**Bruce, et al. 2005**] Bruce, N. D., & Tsotsos, J. K. (2005). Saliency based on information maximization. In *Advances in neural information processing systems* (pp. 155-162).

[**Bruce, et al. 2009**] Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, *9*(3), 5.

[**Burr, et al. 1989**] Burr, D. C., Morrone, M. C., & Spinelli, D. (1989). Evidence for edge and bar detectors in human vision. *Vision Research*, *29*(4), 419-431.

[**Buschman, et al. 2007**] Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *science*, *315*(5820), 1860-1862.

[**Butko, et al. 2008**] Butko, N. J., Zhang, L., Cottrell, G. W., & Movellan, J. R. (2008, May). Visual saliency model for robot cameras. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on* (pp. 2398-2403). IEEE.

[**Buttazzo, et al. 2008**] Buttazzo, G., & Manzotti, R. (2008). Artificial consciousness: Theoretical and practical issues. *Artificial intelligence in medicine*, *44*(2), 79-82.

[**Cakir, et al. 2011**] Cakir, F., Güdükbay, U., & Ulusoy, Ö. (2011). Nearest-Neighbor based Metric Functions for indoor scene recognition. *Computer Vision and*

Image Understanding, *115*(11), 1483-1492.

[**Candes, et al. 2006**] Candes, E. J., Romberg, J. K., & Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, *59*(8), 1207-1223.

[**Caporello Bluvas, et al. 2013**] Caporello Bluvas, E., & Gentner, T. Q. (2013). Attention to natural auditory signals. *Hearing research*, *305*, 10-18.

[**Chachada, et al. 2013**] Chachada, S., & Kuo, C. C. J. (2013, October). Environ -mental sound recognition: A survey. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific* (pp. 1-9). IEEE.

[**Chang, et al. 2010**] Chang, C. K., Siagian, C., & Itti, L. (2010, October). Mobile robot vision navigation & localization using gist and saliency. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on* (pp. 4147-4154). IEEE.

[**Chang, et al. 2011a**] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

[**Chang, et al. 2011b**] Chang, K. Y., Liu, T. L., Chen, H. T., & Lai, S. H. (2011, November). Fusing generic objectness and visual saliency for salient object detection. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 914-921). IEEE.

[**Chatfield, et al. 2011**] Chatfield, K., Lempitsky, V., Vedaldi, A., & Zisserman, A.

(2011). The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of British Machine Vision Conference (BMVC)*, Dundee, UK.

[**Chauhan, et al. 2013**] Chauhan, T. G., Soni, H., & Zafar, S. (2013). A Novel Approach for Text Dependent Speaker Recognition. *International Journal of Research in Computer and Communication Technology (IJRCCT)*, *2*(9), 754-758.

[**Chella, et al. 2007a**] Chella, A., & Manzotti, R. (2007). Artificial consciousness. Exeter: Imprint Academic.

[**Chella, et al. 2007b**] Chella, A., & Manzotti, R. (2007). Artificial intelligence and consciousness. In *Association for the advancement of Artificial Intelligence Fall Symposium* (pp. 1-8).

[**Chen, et al. 1998**] Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, *20*(1), 33-61.

[**Chen, et al. 2014**] Chen, Y., Song, M., Xue, L., Chen, X., & Wang, M. (2014). An audio-visual human attention analysis approach to abrupt change detection in videos. *Signal Processing*. Available online 19 August 2014.

[**Cherry 1953**] Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, *25*(5), 975-979.

[**Cheng, et al. 2011**] Cheng, M. M., Zhang, G. X., Mitra, N. J., Huang, X., & Hu, S. M.

(2011, June). Global contrast based salient region detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 409-416). IEEE.

[**Cheng, et al. 2014**] Cheng, M. M., Zhang, Z., Lin, W. Y., & Torr, P. (2014). BING: Binarized normed gradients for objectness estimation at 300fps. In *IEEE CVPR*.

[**Chia Ai, et al. 2012**] Chia Ai, O., Hariharan, M., Yaacob, S., & Sin Chee, L. (2012). Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems with Applications*, *39*(2), 2157-2165.

[**Chu, et al. 2004**] Chu, S. M., Libal, V., Marcheret, E., Neti, C., & Potamianos, G. (2004, June). Multistage information fusion for audio-visual speech recognition. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on* (Vol. 3, pp. 1651- 1654). IEEE.

[**Chu, et al. 2006**] Chu, S., Narayanan, S., Kuo, C. C., & Mataric, M. J. (2006, July). Where am I? Scene recognition for mobile robots using audio features. In *Multimedia and Expo, 2006 IEEE International Conference on* (pp. 885-888). IEEE.

[**Chu, et al. 2008**] Chu, S., Narayanan, S., & Kuo, C. C. (2008, March). Environmental sound recognition using MP-based features. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* (pp. 1-4). IEEE.

[**Chu, et al. 2009**] Chu, S., Narayanan, S., & Kuo, C. C. (2009). Environmental sound recognition with time-frequency audio features. *Audio, Speech, and*

Language Processing, IEEE Transactions on, *17*(6), 1142-1158.

[**Cortes, et al. 1995**] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[**Cowling, et al. 2003**] Cowling, M., & Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, *24*(15), 2895-2907.

[**Criminisi, et al. 2010**] Criminisi, A., Sharp, T., Rother, C., & Perez, P. (2010). Geodesic image and video editing. *ACM Trans. Graph.*, *29*(5), 134.

[**Cula, et al. 2001**] Cula, O. G., & Dana, K. J. (2001). Compact representation of bidirectional texture functions. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (Vol. 1, pp. 1041-1041). IEEE.

[**Dalal, et al. 2005**] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 1, pp. 886-893). IEEE.

[**Dalal, et al. 2006**] Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Computer Vision-ECCV 2006* (pp. 428-441). Springer Berlin Heidelberg.

[**Daugman 1980**] Daugman, J. G. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. *Vision research*, *20*(10), 847-856.

[**Daugman 1985**] Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, *2*(7), 1160-1169.

[**Davis, et al. 1997**] Davis, G., Mallat, S., & Avellaneda, M. (1997). Adaptive greedy approximations. *Constructive approximation*, *13*(1), 57-98.

[**De Valois, et al. 1982**] De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision research*, *22*(5), 545-559.

[**Dehaene, et al. 2001**] Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1), 1-37.

[**Dehaene, et al. 2006**] Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in cognitive sciences*, *10*(5), 204-211.

[**Dennis, et al. 2011**] Dennis, J., Tran, H. D., & Li, H. (2011). Spectrogram image feature for sound event classification in mismatched conditions. *Signal Processing Letters, IEEE*, *18*(2), 130-133.

[**Dennis, et al. 2013a**] Dennis, J., Tran, H. D., & Chng, E. S. (2013). Overlapping sound event recognition using local spectrogram features and the generalised hough transform. *Pattern Recognition Letters*, *34*(9),1085-1093.

[**Dennis, et al. 2013b**] Dennis, J., Yu, Q., Tang, H., Tran, H. D., & Li, H. (2013, May). Temporal coding of local spectrogram features for robust sound recognition.

In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 803-807). IEEE.

[**Deutsch, et al. 1963**] Deutsch, J. A., & Deutsch, D. (1963). Attention: some theoretical considerations. *Psychological review*, *70*(1), 80.

[**Donoho 2006**] Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal l^1-Norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, *59*(6), 797-829..

[**Donoho, et al. 2008**] Donoho, D. L., & Tsaig, Y. (2008). Fast solution of l^1-norm minimization problems when the solution may be sparse. *Information Theory, IEEE Transactions on*, *54*(11), 4789-4812.

[**Duangudom, et al. 2007**] Duangudom, V., & Anderson, D. V. (2007, September). Using auditory saliency to understand complex auditory scenes. In *15th European Signal Processing Conference (EUSIPCO)* (pp. 1206-1210).

[**Edelman, et al. 2000**] Edelman, G. M., & Tononi, G. (2000). *A universe of consciousness: How matter becomes imagination*. Allen Lane, London.

[**Edelman, et al. 2013**] Edelman, G. M., & Tononi, G. (2013). *Consciousness: How matter becomes imagination*. Penguin UK.

[**Ejaz, et al. 2013**] Ejaz, N., Mehmood, I., & Wook Baik, S. (2013). Efficient visual attention based framework for extracting key frames from videos. *Signal Processing: Image Communication*, *28*(1), 34-44.

[**Elad, et al. 2006a**] Elad, M., & Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, *15*(12), 3736-3745.

[**Elad, et al. 2006b**] Elad, M., & Aharon, M. (2006, June). Image denoising via learned dictionaries and sparse representation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (Vol. 1, pp. 895-900). IEEE.

[**Eriksson, et al. 2008**] Eriksson, J., Larsson, A., & Nyberg, L. (2008). Item-specific training reduces prefrontal cortical involvement in perceptual awareness. *Cognitive Neuroscience, Journal of*, *20*(10), 1777-1787.

[**Evangelopoulos, et al. 2008**] Evangelopoulos, G., Rapantzikos, K., Maragos, P., Avrithis, Y., & Potamianos, A. (2008). Audiovisual attention modeling and salient event detection. In *Multimodal Processing and Interaction* (pp. 1-21). Springer US.

[**Everingham, et al. 2008**] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2008). The PASCAL Visual Object Classes Challenge 2007 (VOC 2007) Results (2007). In *URL* *http://www.pascal-network.org/-challenges/VOC/voc2007/workshop/index. html*.

[**Farinella, et al. 2014**] Farinella, G. M., Ravì, D., Tomaselli, V., Guarnera, M., & Battiato, S. (2014). Representing scenes for real-time context classification on mobile devices. *Pattern Recognition*. Available online 9 June 2014.

[**Feng, et al. 2013**] Feng, K. P., & Yuan, F. (2013, December). Static hand gesture recognition based on HOG characters and support vector machines. In *Instrumentation and Measurement, Sensor Network and Automation*

(IMSNA), 2013 2nd International Symposium on (pp. 936-938). IEEE.

[**Fernández, et al. 2013**] Fernández, A., Álvarez, M. X., & Bianconi, F. (2013). Texture description through histograms of equivalent patterns. *Journal of mathematical imaging and vision*, *45*(1), 76-102

[**Field 1987**] Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, *4*(12), 2379-2394.

[**Field 1994**] Field, D. J. (1994). What is the goal of sensory coding?. *Neural computation*, *6*(4), 559-601.

[**Fisher, et al. 2000**] Fisher , J. W., Darrell, T., Freeman, W. T., & Viola, P. A. (2000, November). Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems (NIPS)*, 13, 772-778.

[**Fisher, et al. 2004**] Fisher, J. W., & Darrell, T. (2004). Speaker association with signal-level audiovisual fusion. *Multimedia, IEEE Transactions on*, *6*(3), 406-413.

[**Friedman 1979**] Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of experimental psychology: General*, *108*(3), 316.

[**Friedman 1996**] Friedman, J. (1996). *Another approach to polychotomous classification* (Vol. 56). Technical report, Department of Statistics, Stanford University.

[**Frintrop, et al. 2010**] Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, *7*(1), 6.

[**Fritz, et al. 2007**] Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention - focusing the searchlight on sound. *Current opinion in neurobiology*, *17*(4), 437-455.

[**Fuertes, et al. 2002**] Fuertes, C. T., & Russ, G. (2002, October). Unification of perception sources for perceptive awareness automatic systems. In *Africon Conference in Africa, 2002. IEEE AFRICON. 6th* (Vol. 1, pp. 283-286). IEEE.

[**Gabor 1946**] Gabor, D. (1946). Theory of communication. *Journal of the Institute of Electrical Engineers*, 93, 429–457.

[**Gao, et al. 2008**] Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). The discriminant center-surround hypothesis for bottom-up saliency. In *Advances in neural information processing systems* (pp. 497-504).

[**Gao, et al. 2009**] Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(6), 989-1005.

[**Ghosal, et al. 2012**] Ghosal, A., Chakraborty, R., Dhara, B. C., & Saha, S. K. (2012, September). Song/instrumental classification using spectrogram based contextual features. In *Proceedings of the CUBE International Information Technology Conference* (pp. 21-25). ACM.

[**Giannakopoulos, et al. 2010**] Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., & Theodoridis, S. (2010). Audio-visual fusion for detecting violent scenes in videos. In *Artificial Intelligence: Theories, Models and Applications* (pp. 91-100). Springer Berlin Heidelberg.

[**Goferman, et al. 2012**] Goferman, S., Zelnik-Manor, L., & Tal, A. (2012). Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(10), 1915-1926.

[**Goldstein 2013**] Goldstein, E. (2013). *Sensation and perception*. Cengage Learning.

[**Gouzoulis-Mayfrank, et al. 2006**] Gouzoulis-Mayfrank, E., Arnold, S., & Heekeren, K. (2006). Deficient inhibition of return in schizophrenia - further evidence from an independent sample. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *30*(1), 42-49.

[**Gray 1984**] Gray, R. M. (1984). Vector quantization. *ASSP Magazine, IEEE*, *1*(2), 4-29.

[**Griffiths, et al. 2003**] Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In *Advances in neural information processing system* (Vol. 15, pp. 11-18). Cambridge, MA: MIT Press.

[**Griffiths, et al. 2004**] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science*, 101, 5228-5235.

[**Griffiths, et al. 2005**] Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005) Integrating topics and syntax. In *Advances in Neural Information*

Processing 17. Cambridge, MA: MIT Press.

[**Guo, et al. 2008**] Guo, C., Ma, Q., & Zhang, L. (2008, June). Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1-8). IEEE.

[**Guo, et al. 2012**] Guo, X., Toyoda, Y., Li, H., Huang, J., Ding, S., & Liu, Y. (2012). Environmental sound recognition using time-frequency intersection patterns. *Applied Computational Intelligence and Soft Computing*, *2012*, 2.

[**Han, et al. 2006**] Han, J., Ngan, K. N., Li, M., & Zhang, H. J. (2006). Unsupervised extraction of visual attention objects in color images. *Circuits and Systems for Video Technology, IEEE Transactions on*, *16*(1), 141-145.

[**Han, et al. 2013a**] Han, J., He, S., Qian, X., Wang, D., Guo, L., & Liu, T. (2013). An object-oriented visual saliency detection framework based on sparse coding representations. *IEEE transactions on circuits and systems for video technology*, *23*(12), 2009-2021.

[**Han, et al. 2013b**] Han, J., Pauwels, E. J., & De Zeeuw, P. (2013). Fast saliency-aware multi-modality image fusion. *Neurocomputing*, *111*, 70-80.

[**Harel, et al. 2006**] Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Advances in neural information processing systems* (pp. 545-552).

[**Hassanieh, et al. 2012**] Hassanieh, H., Indyk, P., Katabi, D., & Price, E. (2012, May). Nearly optimal sparse Fourier transform. In *Proceedings of the forty-fourth* *annual ACM symposium on Theory of computing* (pp. 563-578). ACM.

[**He, et al. 1990**] He, D. C., & Wang, L. (1990). Texture unit, texture spectrum, and texture analysis. *Geoscience and Remote Sensing, IEEE Transactions on*, *28*(4), 509-512.

[**He, et al. 2011**] He, S., Han, J., Hu, X., Xu, M., Guo, L., & Liu, T. (2011, November). A biologically inspired computational model for image saliency detection. In *Proceedings of the 19th ACM international conference on Multimedia* (pp. 1465-1468). ACM.

[**Henriksen, et al. 2007**] Henriksen, J. J. (2007). 3D surface tracking and approximation using Gabor filters. *South Denmark University (March 28, 2007).*

[**Hofmann 2001**] Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning Journal*, 42(1), 177-196.

[**Holland 2003**] Holland, O. (Ed.). (2003). *Machine consciousness*. Exeter: Imprint Academic.

[**Hossan, et al. 2010**] Hossan, M. A., Memon, S., & Gregory, M. A. (2010, December). A novel approach for MFCC feature extraction. In *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on* (pp. 1-5). IEEE.

[**Hou, et al. 2007**] Hou, X., & Zhang, L. (2007, June). Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (pp. 1-8). IEEE.

[**Hoyer 2003**] Hoyer, P. O. (2003). Modeling receptive fields with non-negative sparse coding. *Neurocomputing*, *52*, 547-552.

[**Hoyer, et al. 2002**] Hoyer, P. O., & Hyvärinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision research*, *42*(12), 1593-1605.

[**Hsu, et al. 2002**] Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, *13*(2), 415-425.

[**Huang, et al. 2014**] Huang, Y., Wu, Z., Wang, L., & Song, C. (2014). Multiple spatial pooling for visual object recognition. *Neurocomputing*, *129*, 225-231.

[**Hubel, et al. 1968**] Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, *195*(1), 215-243.

[**Itti, et al. 2001**] Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, *2*(3), 194-203.

[**Itti, et al. 1998**] Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254-1259.

[**Itti 2007**] Itti, L. (2007). Visual salience. *Scholarpedia*, *2*(9), 3327.

[**Janvier, et al. 2012**] Janvier, M., Alameda-Pineda, X., Girinz, L., & Horaud, R. (2012, November). Sound-event recognition with a companion humanoid. In *Humanoid Robots (Humanoids), 2012 12th IEEE-RAS International Conference on* (pp. 104-111). IEEE.

[**Jégou, et al. 2011**] Jégou, H., Douze, M., & Schmid, C. (2011). Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(1), 117-128.

[**Jennings 2000**] Jennings, C. (2000). In search of consciousness. *Nature Neuroscience*, *3*(8), 1.

[**Ji, et al. 2013**] Ji, Z., Wang, J., Su, Y., Song, Z., & Xing, S. (2013). Balance between object and background: Object-enhanced features for scene image classification. *Neurocomputing*, 120, 15-23.

[**Jones, et al. 1987**] Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, *58*(6), 1233-1258.

[**Kahneman 1973**] Kahneman, D. (1973). *Attention and effort* (p. 246). Englewood Cliffs, NJ: Prentice-Hall.

[**Kalinli, et al. 2007**] Kalinli, O., & Narayanan, S. S. (2007, August). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In Proceedings of *INTERSPEECH* (pp. 1941-1944).

[**Kalinli, et al. 2009**] Kalinli, O., Sundaram, S., & Narayanan, S. (2009, October).

Saliency-driven unstructured acoustic scene classification using latent perceptual indexing. In *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on* (pp. 1-6). IEEE.

[**Kanwisher 2001**] Kanwisher, N. (2001). Neural events and perceptual awareness. *Cognition*, *79*(1), 89-113.

[**Karbasi, et al. 2011**] Karbasi, M., Ahadi, S. M., & Bahmanian, M. (2011, December). Environmental sound classification using spectral dynamic features. In*Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on* (pp. 1-5). IEEE.

[**Kaya, et al. 2012**] Kaya, E. M., & Elhilali, M. (2012, March). A temporal saliency map for modeling auditory attention. In *Information Sciences and Systems (CISS), 2012 46th Annual Conference on* (pp. 1-6). IEEE.

[**Kayser, et al. 2005**] Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, *15*(21), 1943-1947.

[**Khunarsal, et al. 2013**] Khunarsal, P., Lursinsap, C., & Raicharoen, T. (2013). Very short time environmental sound classification based on spectrogram pattern matching. *Information Sciences*, *243*, 57-74.

[**Kim, et al. 2013**] Kim, B., Ban, S. W., & Lee, M. (2013). Top-down attention based on object representation and incremental memory for knowledge building and inference. *Neural Networks*, *46*, 9-22.

[**Kim, et al. 2014**] Kim, K., Lin, K. H., Walther, D. B., Hasegawa-Johnson, M. A., &

Huang, T. S. (2014). Automatic detection of auditory salience with optimized linear filters derived from human annotation. *Pattern Recognition Letters*, *38*, 78-85.

[**Klein, et al. 2011**] Klein, D. A., & Frintrop, S. (2011, November). Center-surround divergence of feature statistics for salient object detection. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 2214-2219). IEEE.

[**Knerr, et al. 1990**] Knerr, S., Personnaz, L., & Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing* (pp. 41-50). Springer Berlin Heidelberg.

[**Ko, et al. 2006**] Ko, B. C., & Nam, J. Y. (2006). Object-of-interest image segmentation based on human attention and semantic region clustering. *JOSA A*, *23*(10), 2462-2470.

[**Kobayashi, et al. 2014**] Kobayashi, T., & Ye, J. (2014, May). Acoustic feature extraction by statistics based local binary pattern for environmental sound classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 3052-3056). IEEE.

[**Koch, et al. 1985**] Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, *4*(4), 219-227.

[**Koivisto, et al. 2009**] Koivisto, M., Kainulainen, P., & Revonsuo, A. (2009). The relationship between awareness and attention: evidence from ERP responses. *Neuropsychologia*, *47*(13), 2891-2899.

[**Kovashka, et al. 2012**] Kovashka, A., Parikh, D., & Grauman, K. (2012, June). Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 2973-2980). IEEE.

[**Kuipers 2005**] Kuipers, B. (2005, July). Consciousness: Drinking from the firehose of experience. In *AAAI* (pp. 1298-1305).

[**Lades, et al. 1993**] Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R. P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *Computers, IEEE Transactions on*, *42*(3), 300-311.

[**Lai, et al. 2012**] Lai, J. L., & Yi, Y. (2012). Key frame extraction based on visual attention model. *Journal of Visual Communication and Image Representation*, *23*(1), 114-125.

[**Lallemand, et al. 2012**] Lallemand, I., Schwarz, D., & Artières, T. (2012). Content-based retrieval of environmental sounds by multiresolution analysis. *Sound and Music Computing*, Copenhagen, Denmark.

[**Lampert, et al. 2009**] Lampert, C. H., Blaschko, M. B., & Hofmann, T. (2009). Efficient subwindow search: A branch and bound framework for object localization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(12), 2129-2142.

[**Lee, et al. 1999**] Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature neuroscience*, *2*(4), 375-381.

[**Lee, et al. 2006**] Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2006). Efficient sparse coding algorithms. In *Advances in neural information processing systems* (pp. 801-808).

[**Lefter, et al. 2013**] Lefter, I., Rothkrantz, L. J., & Burghouts, G. J. (2013). A comparative study on automatic audio–visual fusion for aggression detection using meta-information. *Pattern Recognition Letters*, *34*(15), 1953-1963.

[**Li, et al. 2008**] Li, X., Tao, D., Maybank, S. J., & Yuan, Y. (2008). Visual music and musical vision. *Neurocomputing*, *71*(10), 2023-2028.

[**Li, et al. 2010**] Li, Y., & Li, Y. (2010, December). Eco-environmental sound classification based on matching pursuit and support vector Machine. In *information engineering and computer science (ICIECS), 2010 2nd international conference on* (pp. 1-4). IEEE.

[**Li, et al. 2011a**] Li, J., Levine, M., An, X. J., & He, H. E. (2011). Saliency detection based on frequency and spatial domain analyses. In *Proceedings of the British Machine Vision Conference* (pp. 86.1-86.11). BMVC.

[**Li, et al. 2011b**] Li, Q., Zhou, Y., & Yang, J. (2011, July). Saliency based image segmentation. In *Multimedia Technology (ICMT), 2011 International Conference on* (pp. 5068-5071). IEEE.

[**Li, et al. 2012**] Li, Z., Herfet, T., Grochulla, M., & Thormahlen, T. (2012, September). Multiple active speaker localization based on audio-visual fusion in two stages. In *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on* (pp. 1-7). IEEE.

[**Li, et al. 2013**] Li, S. Z., Huang, H. F., & Xiao, N. F. (2013). MapReduce implementation of face recognition based on HOG feature and NSMD. *Applied Mechanics and Materials*, *385*, 1572-1575.

[**Liang, et al. 2013**] Liang, J., & Yuen, S. Y. (2013). An edge detection with automatic scale selection approach to improve coherent visual attention model. *Pattern Recognition Letters*, *34*(13), 1519-1524.

[**Lin, et al. 2012**] Lin, K. H., Zhuang, X., Goudeseune, C., King, S., Hasegawa-Johnson, M., & Huang, T. S. (2012, March). Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on* (pp. 2277-2280). IEEE.

[**Liu, et al. 2002**] Liu, C., & Wechsler, H. (2002). Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *Image processing, IEEE Transactions on*, *11*(4), 467-476.

[**Liu, et al. 2009**] Liu, C. C., Doong, J. L., Hsu, C. C., Huang, W. S., & Jeng, M. C. (2009). Evidence for the selective attention mechanism and dual-task interference. *Applied ergonomics*, *40*(3), 341-347.

[**Liu, et al. 2014a**]Liu, J., Huang, Y., Wang, L., & Wu, S. (2014). Hierarchical feature coding for image classification. *Neurocomputing*, 144, 509-515.

[**Liu, et al. 2014b**] Liu, Y., Yang, Y., Xie, H., & Tang, S. (2014). Fusing audio vocabulary with visual features for pornographic video detection. *Future Generation Computer Systems*, *31*, 69-76.

[**Liu, et al. 2013**] Liu, H., Engelke, U., Wang, J., Le Callet, P., & Heynderickx, I. (2013). How does image content affect the added value of visual attention in objective image quality assessment?. *IEEE Signal Processing Letters*, *20*(4), 355-358.

[**Lloyd 1982**] Lloyd, S. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, *28*(2), 129-137.

[**Lowe 1999**] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* (Vol. 2, pp. 1150-1157). IEEE.

[**Lowe 2004**] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91-110.

[**Ma, et al. 2003**] Ma, Y. F., & Zhang, H. J. (2003, November). Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the eleventh ACM international conference on Multimedia* (pp. 374-381). ACM.

[**MacPherson, et al. 2003**] MacPherson, A. C., Klein, R. M., & Moore, C. (2003). Inhibition of return in children and adolescents. *Journal of Experimental Child Psychology*, *85*(4), 337-351.

[**Mahadevan, et al. 2010**] Mahadevan, V., & Vasconcelos, N. (2010). Spatiotemporal saliency in dynamic scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*(1), 171-177.

[**Mairal, et al. 2008**] Mairal, J., Elad, M., & Sapiro, G. (2008). Sparse representation for color image restoration. *Image Processing, IEEE Transactions on*, *17*(1), 53-69.

[**Majozi, et al. 2005**] Majozi, T., & Zhu, X. X. (2005). A combined fuzzy set theory and MILP approach in integration of planning and scheduling of batch plants - Personnel evaluation and allocation. *Computers and Chemical Engineering*, *29*, 2029-2047.

[**Mallat, et al. 1993**] Mallat, S. G., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, *41*(12), 3397-3415.

[**Manzotti 2006**] Manzotti, R. (2006). An alternative view of conscious perception. *Journal of Consciousness Studies*, *13*(6), 45-79.

[**Manzotti 2008**] Manzotti, R., & Tagliasco, V. (2008). Artificial consciousness: A discipline between technological and theoretical obstacles. *Artificial intelligence in medicine*, *44*(2), 105-117.

[**Marčelja 1980**] Marčelja, S. (1980). Mathematical description of the responses of simple cortical cells*. *JOSA*, *70*(11), 1297-1300.

[**Marfil, et al. 2009**] Marfil, R., Núñez, P., Bandera, A., & Sandoval, F. (2009). A novel approach for salient image regions detection and description. *Pattern Recognition Letters*, *30*(16), 1464-1476.

[**McDermott 2007**] McDermott, D. (2007). Artificial intelligence and consciousness.

The Cambridge handbook of consciousness, 117-150.

[**Mesgarani, et al. 2012**] Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233-236.

[**Michael, et al. 2011**] Michael, G. A., & Gálvez-García, G. (2011). Salience-based progression of visual attention. *Behavioural brain research*, *224*(1), 87-99.

[**Mitrović, et al. 2010**] Mitrović, D., Zeppelzauer, M., & Breiteneder, C. (2010). Features for content-based audio retrieval. *Advances in computers*, *78*, 71-150.

[**Mondal, et al. 2012**] Modal, S., Pujari S., & Sangiri, T. (2012). Environmental natural sound detection and classification using content-based retrieval (CBR) and MFCC. *International Journal of Engineering Research and Applications (IJERA)*, 2(6), 123-129.

[**Mukherjee, et al. 2013**] Mukherjee, R., Islam, T., & Sankar, R. (2013, April). Text dependent speaker recognition using shifted MFCC. In *Southeastcon, 2013 Proceedings of IEEE* (pp. 1-4). IEEE.

[**Murty, et al. 2006**] Murty, K. S. R., & Yegnanarayana, B. (2006). Combining evidence from residual phase and MFCC features for speaker recognition. *Signal Processing Letters, IEEE*, *13*(1), 52-55.

[**Nakagawa, et al. 2012**] Nakagawa, S., Wang, L., & Ohtsuka, S. (2012). Speaker identification and verification by combining MFCC and phase information. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(4), 1085-1095.

[**Nanni, et al. 2013**] Nanni, L., & Lumini, A. (2013). Heterogeneous bag-of-features for object/scene recognition. *Applied Soft Computing*, *13*(4), 2171-2178.

[**Nirjon, et al. 2013**] Nirjon, S., Dickerson, R., Stankovic, J., Shen, G., & Jiang, X. (2013, February). sMFCC: exploiting sparseness in speech for fast acoustic feature extraction on mobile devices - a feasibility study. In *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications* (p. 8). ACM.

[**Ojala, et al. 2002**] Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(7), 971-987.

[**Oliva, et al. 2001**] Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, *42*(3), 145-175.

[**Olshausen 1996**] Olshausen, B. A. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607-609.

[**Olshausen, et al. 1997**] Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1?. *Vision research*, *37*(23), 3311-3325.

[**Olshausen, et al. 2000**] Olshausen, B. A., & Field, D. J. (2000). Vision and the coding of natural images. *American Scientist*, *88*(3), 238-245.

[**Olshausen 2003**] Olshausen, B. A. (2003). Principles of image representation in visual cortex. *The visual neurosciences*, 1603-1615.

[**Olshausen 2004**] Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. Current *opinion in neurobiology*, *14*(4), 481-487.

[**Palmer, et al. 1981**] Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspect-ive and the perception of objects. *Attention and performance IX*, *1*, 4.

[**Parker, et al. 1988**] Parker, A. J., & Hawken, M. J. (1988). Two-dimensional spatial structure of receptive fields in monkey striate cortex. *JOSA A*, *5*(4), 598-605.

[**Pati, et al. 1993**] Pati, Y. C., Rezaiifar, R., & Krishnaprasad, P. S. (1993, November). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on* (pp. 40-44). IEEE.

[**Perazzi, et al. 2012**] Perazzi, F., Krahenbuhl, P., Pritch, Y., & Hornung, A. (2012, June). Saliency filters: Contrast based filtering for salient region detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 733-740). IEEE.

[**Pérez Grassi, et al. 2011**] Pérez Grassi, A., Frolov, V., & Puente León, F. (2011). Information fusion to detect and classify pedestrians using invariant features. *Information fusion*,*12*(4), 284-292.

[**Petit, et al. 2013**] Petit, C., El-Amraoui, A., & Avan, P. (2013). Audition: hearing and deafness. *Neuroscience in the 21st Century: From Basic to Clinical*, 675-741.

[**Pietikäinen, et al. 2011**] Pietikäinen, M., Hadid, A., Zhao, G., & Ahonen, T. (2011). *Computer vision using local binary patterns* (Vol. 40). Springer.

[**Posner, et al. 1980**] Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of experimental psychology: General*, *109*(2), 160.

[**Raj, et al. 2007**] Raj, B., Smaragdis, P., Shashanka, M., & Singh, R. (2007, January). Separating a foreground singer from background music. In *International Symposium on Frontiers of Research on Speech and Music, Mysore, India*.

[**Ramík, et al. 2013a**] Ramík, D. M., Madani, K., & Sabourin, C. (2013). From visual patterns to semantic description: A cognitive approach using artificial curiosity as the foundation. *Pattern Recognition Letters*, *34*(14), 1577-1588.

[**Ramík, et al. 2013b**] Ramík, D. M., Sabourin, C., & Madani, K. (2013). Autonomous knowledge acquisition based on artificial curiosity: Application to mobile robots in an indoor environment. *Robotics and Autonomous Systems*, *61*(12), 1680-1695.

[**Reggia 2013**] Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112-131.

[**Richards 2000**] Richards, J. E. (2000). Localizing the development of covert attention in infants with scalp event-related potentials. *Developmental Psychology*, *36*(1), 91.

[**Rubinstein, et al. 2010**] Rubinstein, R., Zibulevsky, M., & Elad, M. (2010). Double sparsity: Learning sparse dictionaries for sparse signal approximation. *Signal Processing, IEEE Transactions on*, *58*(3), 1553-1564.

[**Ruesch, et al. 2008**] Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., & Pfeifer, R. (2008, May). Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on* (pp. 962-967). IEEE.

[**Rutishauser, et al. 2004**] Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004, July). Is bottom-up attention useful for object recognition?. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on* (Vol. 2, pp. II-37). IEEE.

[**Satpathy, et al. 2014**] Satpathy, A., Jiang, X., & Eng, H. L. (2014). LBP-based edge-texture features for object recognition. *IEEE Transactions on Image Processing*, 23, 1953-1964.

[**Scharfenberger, et al. 2013**] Scharfenberger, C., Waslander, S. L., Zelek, J. S., & Clausi, D. A. (2013, May). Existence detection of objects in images for robot vision using saliency histogram features. In *Computer and Robot Vision (CRV), 2013 International Conference on* (pp. 75-82). IEEE.

[**Seth 2009**] Seth, A. (2009). The strength of weak artificial consciousness. *International Journal of Machine Consciousness*, *1*(01), 71-82.

[**Shannon 1948**] Shannon, C. E. (1948). Bell System Tech. J. 27 (1948) 379; CE Shannon. *Bell System Tech. J*, *27*, 623.

[**Shannon 2001**] Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3-55.

[**Simoncelli, et al. 2001**] Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, *24*(1), 1193-1216.

[**Sitte, et al. 2007**] Sitte, R., & Willets, L. (2007, February). Non-speech environmental sound identification for surveillance using self-organizing-maps. In *Proceedings of the Fourth conference on IASTED International Conference: Signal Processing, Pattern Recognition, and Applications* (pp. 281-286). ACTA Press.

[**Sivasankaran, et al. 2013**] Sivasankaran, S., & Prabhu, K. M. M. (2013, January). Robust features for environmental sound classification. In *Electronics, Computing and Communication Technologies (CONECCT), 2013 IEEE International Conference on* (pp. 1-6). IEEE.

[**Sivic, et al. 2003**] Sivic, J., & Zisserman, A. (2003, October). Video Google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (pp. 1470-1477). IEEE.

[**Sivic, et al. 2006**] Sivic, J., & Zisserman, A. (2006). Video Google: Efficient visual search of videos. In *Toward Category-Level Object Recognition* (pp. 127-144). Springer Berlin Heidelberg.

[**Souli, et al. 2011**] Souli, S., & Lachiri, Z. (2011). Environmental sounds classification based on visual features. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (pp. 459-466). Springer Berlin Heidelberg.

[**Souli, et al. 2012a**] Souli, S., & Lachiri, Z. (2012, October). Environmental sound classification using log-Gabor filter. In *Signal Processing (ICSP), 2012 IEEE 11th International Conference on* (Vol. 1, pp. 144-147). IEEE.

[**Souli, et al. 2012b**] Souli, S., & Lachiri, Z. (2012). Environmental sounds spectrogram classification using log-Gabor filters and multiclass support vector machines. *International Journal of Computer Science Issues (IJCSI)*, *9*(4).

[**Spampinato, et al. 2012**] Spampinato, C., & Palazzo, S. (2012, November). Enhancing object detection performance by integrating motion objectness and perceptual organization. In *Pattern Recognition (ICPR), 2012 21st International Conference on* (pp. 3640-3643). IEEE.

[**Stevens, et al. 2012**] Stevens, C., & Bavelier, D. (2012). The role of selective attention on academic foundations: a cognitive neuroscience perspective. *Developmental cognitive neuroscience*, *2*, S30-S48.

[**Stevenson 2010**] Stevenson, A. (Ed.). (2010). *Oxford dictionary of English*. Oxford University Press.

[**Steyvers, et al. 2007**] Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424-440.

[**Su, et al. 2011**] Su, F., Yang, L., Lu, T., & Wang, G. (2011, November). Environmental sound classification for scene recognition using local discriminant bases and HMM. In *Proceedings of the 19th ACM international conference on Multimedia* (pp. 1389-1392). ACM.

[**Sujatha, et al. 2012**] Sujatha, K. S., Keerthana, P., Priya, S. S., Kaavya, E., & Vinod, B. (2012). Fuzzy based multiple dictionary bag of words for image classification. *Procedia Engineering*, *38*, 2196-2206.

[**Sun, et al. 2010**] Sun, X., Yao, H., Ji, R., Xu, P., Liu, X., & Liu, S. (2010, September). Saliency detection based on short-term sparse representation. In *Image Processing (ICIP), 2010 17th IEEE International Conference on* (pp. 1101-1104). IEEE.

[**Tibshirani 1996**] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

[**Treisman 1960**] Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, *12*(4), 242-248.

[**Treisman, et al. 1980**] Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

[**Tsuchida, et al. 2012**] Tsuchida, T., & Cottrell, G. W. (2012). Auditory saliency using natural statistics. In *Annual meeting of the cognitive science society* (pp.

1048-1053).

[**Uzkent, et al. 2012**] Uzkent, B., Barkana, B. D., & Cevikalp, H. (2012). Non-speech environmental sound classification using SVMs with a new set of features. *International Journal of Innovative Computing, Information and Control*, *8*(5B), 3511-3524.

[**Vinje, et al. 2000**] Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*(5456), 1273-1276.

[**Walther, et al. 2005**] Walther, D., Rutishauser, U., Koch, C., & Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, *100*(1), 41-63.

[**Wang, et al. 2004**] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, *13*(4), 600-612.

[**Wang, et al. 2008**] Wang, J. C., Lee, H. P., Wang, J. F., & Lin, C. B. (2008). Robust environmental sound recognition for home automation. *Automation Science and Engineering, IEEE Transactions on*, *5*(1), 25-31.

[**Wang, et al. 2010a**] Wang, W., Barnaghi, P., & Bargiela, A. (2010). Probabilistic topic models for learning terminological ontologies. *Knowledge and Data Engineering, IEEE Transactions on*, *22*(7), 1028-1040.

[**Wang, et al. 2010b**] Wang, W., Wang, Y., Huang, Q., & Gao, W. (2010, June).

Measuring visual saliency by site entropy rate. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (pp. 2368-2375). IEEE.

[**Wang, et al. 2013a**] Wang, J. Y., Zhang, K., Madani, K., & Sabourin, C. (2013, November). Heterogeneous information saliency features' fusion approach for machine's environment sounds based awareness. In *Awareness Science and Technology and Ubi-Media Computing (iCAST-UMEDIA), 2013 International Joint Conference on* (pp. 197-205). IEEE.

[**Wang, et al. 2013b**] Wang, Q., Yan, P., Yuan, Y., & Li, X. (2013). Multi-spectral saliency detection. *Pattern Recognition Letters*, *34*(1), 34-41.

[**Wang, et al. 2014**] Wang, Q., Zhu, G., & Yuan, Y. (2014). Statistical quantization for similarity search. *Computer Vision and Image Understanding*, *124*, 22-30.

[**Wickens, et al. 1990**] Wickens, C. D., & Andre, A. D. (1990). Proximity compatibility and information display: Effects of color, space, and objectness on information integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *32*(1), 61-77.

[**Widmer, et al. 2005**] Widmer, G., Dixon, S., Knees, P., Pampalk, E., & Pohle, T. (2005). From sound to "sense" via feature extraction and machine learning: Deriving high-level descriptors for characterising music. *Sound to Sense: Sense to Sound: A State-of-the-Art*, 161-194.

[**Wright, et al. 2009**] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(2), 210-227.

[**Xu, et al. 2013]** Xu, L., Li, H., Zeng, L., & Ngan, K. N. (2013). Saliency detection using joint spatial-color constraint and multi-scale segmentation. *Journal of Visual Communication and Image Representation*, *24*(4), 465-476.

[**Xu, et al. 2014**] Xu, J., & Yue, S. (2014). Mimicking visual searching with integrated top down cues and low-level features. *Neurocomputing*, *133*, 1-17.

[**Yamakawa, et al. 2011**] Yamakawa, N., Takahashi, T., Kitahara, T., Ogata, T., & Okuno, H. G. (2011). Environmental sound recognition for robot audition using matching-pursuit. In *Modern Approaches in Applied Intelligence* (pp. 1-10). Springer Berlin Heidelberg.

[**Yang, et al. 2009**] Yang, J. C., Yu, K., Gong, Y. H., & Huang, T. (2009, June). Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on* (pp. 1794-1801). IEEE.

[**Yang, et al. 2010**] Yang, M., & Zhang, L. (2010). Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary. In *Computer Vision–ECCV 2010* (pp. 448-461). Springer Berlin Heidelberg.

[**Yang, et al. 2013**] Yang, J., Lin, T., & Jin, X. (2013). An Image Sparse Representation for Saliency Detection. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, *11*(10), 6143-6150.

[**Yantis, et al. 2003**] Yantis, S., & Serences, J. T. (2003). Cortical mechanisms of space-based and object-based attentional control. *Current opinion in neurobiology*, *13*(2), 187-193.

[**Yanulevskaya, et al. 2013**] Yanulevskaya, V., Uijlings, J., & Geusebroek, J. M. (2013). Salient object detection: From pixels to segments. *Image and Vision Computing*, *31*(1), 31-42.

[**Yarbus 1967**] Yarbus, A. L. (1967). *Eye movements and vision* (Vol. 2, No. 5.10). L. A. Rigss (Ed.). New York: Plenum press.

[**Yegnanarayana, et al. 2002**] Yegnanarayana, B., & Kishore, S. P. (2002). AANN: an alternative to GMM for pattern recognition. *Neural Networks*, *15*(3), 459-469.

[**Yeh, et al. 2014**] Yeh, H. H., Liu, K. H., & Chen, C. S. (2014). Salient object detection via local saliency estimation and global homogeneity refinement. *Pattern Recognition*, *47*(4), 1740-1750.

[**Yu, et al. 2008**] Yu, G., & Slotine, J. J. (2008, December). Fastwavelet-based visual classification. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on* (pp. 1-5). IEEE.

[**Yu, et al. 2009**] Yu, G., & Slotine, J. J. (2009, April). Audio classification from time-frequency texture. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (pp. 1677-1680). IEEE.

[**Zadeh 1965**] Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338-353.

[**Zeng, et al. 2010**] Zeng, C., & Ma, H. (2010, August). Robust head-shoulder detection by PCA-based multilevel HOG-LBP detector for people counting.

In *Pattern Recognition (ICPR), 2010 20th International Conference on* (pp. 2069-2072). IEEE.

[**Zhang, et al. 2008**] Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of vision*, *8*(7), 32.

[**Zhang, et al. 2011**] Zhang, Y., Ogata, T., Nishide, S., Takahashi, T., & Okuno, H. G. (2011). Classification of known and unknown environmental sounds based on self-organized space using a recurrent neural network. *Advanced Robotics*, *25*(17), 2127-2141.

[**Zhang, et al. 2013**] Zhang, X., & Li, Y. (2013). Environmental sound recognition using double-level energy detection. *Journal of Signal and Information Processing*, *4*, 19.

[**Zhang, et al. 2013**] Zhang, C., Wang, S., Huang, Q., Liu, J., Liang, C., & Tian, Q. (2013). Image classification using spatial pyramid robust sparse coding. *Pattern Recognition Letters*, *34*(9), 1046-1052.

[**Zhang, et al. 2014**] Zhang, X., & Zhang, X. (2014). Salient region detection based on global contrast and object-biased Gaussian refinement. *Journal of Multimedia*, *9*(7), 941-947.

[**Zhang, et al. 2014**] Zhang, D., & Liu, C. (2014). A salient object detection framework beyond top-down and bottom-up mechanism. *Biologically Inspired Cognitive Architectures*. Available online 21 July 2014

[**Zhao, et al. 2011**] Zhao, C., Liu, C., & Lai, Z. (2011). Multi-scale gist feature manifold for building recognition. *Neurocomputing*, *74*(17), 2929-2940.

[**Zhao, et al. 2013**] Zhao, Q., & Koch, C. (2013). Learning saliency-based visual attention: A review. *Signal Processing*, *93*(6), 1401-1407.

[**Zhou, et al. 2014**] Zhou, C., & Liu, C. (2014). Semantic image segmentation using low-level features and contextual cues. *Computers & Electrical Engineering*, *40*(3), 844-857.

[**Zimmermann 2010**] Zimmermann, H. J. (2010). Fuzzy set theory. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 317-332.

[**Zlatintsi, et al. 2012**] Zlatintsi, A., Maragos, P., Potamianos, A., & Evangelopoulos, G. (2012, August). A saliency-based approach to audio event detection and summarization. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European* (pp. 1294-1298). IEEE.

[**Zubair, et al. 2013**] Zubair, S., Yan, F., & Wang, W. (2013). Dictionary learning based sparse coefficients for audio classification with max and average pooling. *Digital Signal Processing*, *23*(3), 960-970.

Résumé :

L’objectif principal de cette thèse porte sur la conception d’un système de perception artificiel permettant d’identifier des scènes ou évènements pertinents dans des environnements complexes. Les travaux réalisés ont permis d’étudier et de mettre en œuvre d’un système de perception bio-inspiré basé sur l’attention visuelle et auditive. Les principales contributions de cette thèse concernent la saillance auditive associée à une identification des sons et bruits environnementaux ainsi que la saillance visuelle associée à une reconnaissance d’objets pertinents. La saillance du signal sonore est calculée en fusionnant des informations extraites des représentations temporelles et spectrales du signal acoustique avec une carte de saillance visuelle du spectrogramme du signal concerné. Le système de perception visuelle est quant à lui composé de deux mécanismes distincts. Le premier se base sur des méthodes de saillance visuelle et le deuxième permet d’identifier l’objet en premier plan. D’autre part, l’originalité de notre approche est qu’elle permet d’évaluer la cohérence des observations en fusionnant les informations extraites des signaux auditifs et visuels perçus. Les résultats expérimentaux ont permis de confirmer l’intérêt des méthodes utilisées dans le cadre de l’identification de scènes pertinentes dans un environnement complexe.

Abstract:

The main goal of these researches is the design of one artificial perception system allowing to identify events or scenes in a complex environment. The work carried out during this thesis focused on the study and the conception of a bio-inspired perception system based on the both visual and auditory saliency. The main contributions of this thesis are auditory saliency with sound recognition and visual saliency with object recognition. The auditory saliency is computed by merging information from the both temporal and spectral signals with a saliency map of a spectrogram. The visual perception system is based on visual saliency and recognition of foreground object. In addition, the originality of the proposed approach is the possibility to do an evaluation of the coherence between visual and auditory observations using the obtained information from the features extracted from both visual and auditory patters. The experimental results have proven the interest of this method in the framework of scene identification in a complex environment.