8-2016

Face Centered Image Analysis Using Saliency and Deep Learning Based Techniques

Rui Guo
University of Tennessee, Knoxville, rguo1@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Rui Guo entitled "Face Centered Image Analysis Using Saliency and Deep Learning Based Techniques." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Computer Engineering.

Hairong Qi, Major Professor

We have read this dissertation and recommend its acceptance:

Jens Gregor, Lynne Parker, Yulong Xing

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Face Centered Image Analysis Using Saliency and Deep Learning Based Techniques

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Rui Guo

August 2016

To my parents, my wife and lovely daughter

Acknowledgements

The past 5 years are full of memories with smiles and tears. Though only my name appears on the cover of the dissertation, it is definitely a great many people have contributed to its generation. I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my advisor, Dr. Hairong Qi. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. Dr. Qi taught me how to question thoughts and express ideas. Her patience and support helped me overcome many crisis situations and finish this dissertation. I am always lucky to have her around not only input her time to my research but always willing to guide me in the life. My childish and temper are always tolerant in her good heart. The lessons I learnt would benefit me here and forever. Meanwhile, I would like to thank my committee members, Dr. Jens Gregor, Dr. Lynn Parker and Dr. Yulong Xing. Their insights about research inspire me in many ways. I really appreciate the time they input to my whole dissertation and defense including how to name a research work. I learnt a lot from their broad view in the frontier of machine learning, computer vision and algorithms. The spirits of rigid learning have been planted in my heart.

The outcome of my current learning has a strong tie to the environment which is built in our lab. I have lots of unforgotten days spending with my colleagues Li He, Jiajia Luo, Zhibo Wang, Shuangjiang Li, Liu Liu, Austin Albright, Daniel Capilla,

Bryan Bodkin, Yang Song, Zhifei Zhang, Alireza Rahimpour, Ali Taalimi, Ying Qu and Chengcheng Li. Not only the supports and courage you gave to me, but also the friendship built in our group will be memorized. Once an AICIPer, always an AICIPer!

Last but not least, I would like to express my deepest appreciation to my parents and my wife, for their unconditional support and encouragement. Their dedication and love are always the biggest motivation for any of my achievements. They are the source of the continuous power to make me fearless in the endeavor. My lovely daughter Evelyn is my greatest achievement. Love you all forever!

Abstract

Image analysis starts with the purpose of configuring vision machines that can perceive like human to intelligently infer general principles and sense the surrounding situations from imagery. This dissertation studies the face centered image analysis as the core problem in high level computer vision research and addresses the problem by tackling three challenging subjects: Are there anything interesting in the image? If there is, what is/are that/they? If there is a person presenting, who is he/she? What kind of expression he/she is performing? Can we know his/her age? Answering these problems results in the saliency-based object detection, deep learning structured objects categorization and recognition, human facial landmark detection and multitask biometrics.

To implement object detection, a three-level saliency detection based on the self-similarity technique (SMAP) is firstly proposed in the work. The first level of SMAP accommodates statistical methods to generate proto-background patches, followed by the second level that implements local contrast computation based on image self-similarity characteristics. At last, the spatial color distribution constraint is considered to realize the saliency detection. The outcome of the algorithm is a full resolution image with highlighted saliency objects and well-defined edges.

In object recognition, the Adaptive Deconvolution Network (ADN) is implemented to categorize the objects extracted from saliency detection. To improve the system performance, $L1/2$ norm regularized ADN has been proposed and tested in different

applications. The results demonstrate the efficiency and significance of the new structure.

To fully understand the facial biometrics related activity contained in the image, the low rank matrix decomposition is introduced to help locate the landmark points on the face images. The natural extension of this work is beneficial in human facial expression recognition and facial feature parsing research.

To facilitate the understanding of the detected facial image, the automatic facial image analysis becomes essential. We present a novel deeply learnt tree-structured face representation to uniformly model the human face with different semantic meanings. We show that the proposed feature yields unified representation in multi-task facial biometrics and the multi-task learning framework is applicable to many other computer vision tasks.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Reasoning the surrounding environment and analyzing the situation from visual input is not only a fundamental function of human vision system but also a long-term striving goal of Artificial Intelligence and Computer Vision research. The implementation of this ambitious goal would greatly enhance the development in security surveillance and monitoring Thirde et al. (2006), robots visual navigation and planning Newman et al. (2006), abnormal event awareness in large scale social activity Thirde et al. (2006), computer-assisted medical image analysis for disease diagnosis Mirota et al. (2009) and Internet image-based content search and acquisition engine design Li et al. (2009). Each of these applications requires a complicated reasoning in the image domain to distinguish the objects, background regions, geometric positions, color distributions, lighting, 3D structure and their correlated relationship. In the specific computer vision area, the image analysis is referred as to name the scene and objects located in the image. However, this over-simplified answer involves more challenges rather than a completed explanation. We are always pursuing to reach a higher level which is reasoning more semantic properties and structures from the image to enable the deep understanding of the objects, persons and their identities, expression and potential activities.

It is obvious that image analysis has multiple level requirements. Taking human vision generation processing as an example, in a short glance, human can rapidly locate the salient things in the whole perceptive field. After the locating, the refine process starts to recognize the attributes of the sensed objects and then estimate the activity and situation associated with objects. The whole process involves object detection, recognition and high level situation estimation. The straightforward assumption to tackle the long-term goal in image analysis is to decompose the long-term challenge into couple of highly correlated sub-tasks: object detection, recognition and activity analysis. Because of the huge spaces that the information spanned, the decomposition is quite reasonable to simplify the problem and meanwhile keeps it focus on the essential key points to answer the questions: is there anything interesting contained in the image? What are they? Who are they? And what are these people's current status and potential activities? The completed answers to these questions spontaneously formulate the hierarchical procedure of human vision and neutron system corporately processing the visual input and so forth implementing the entire perceptive mechanism.

1.1 Saliency Based Object Detection

Human visual system has an incredible capability to implement the focus of attention mechanism. This judgment capability enables the visual system to rapidly and efficiently filter the important regions or objects out of the surrounding environment. Related research Grossberg (1995); Treisman and Gelade (1980) reveals that the behavior of the visual system is guided by both discriminant analysis and stimulusdriven process. Generally speaking, the global discriminant analysis is related to the human cognitive capability which is a learning process in memory and the neuron system. Millions of special patterns are learned and accumulated from personal experiences. Then classifiers formulated statistically are performed to locate the salient object from its surroundings. Comparatively, the local stimulus-driven

process only asks for short-term, small region response on the image Cheng et al. (2015). Thus, the local contrast becomes the essential factor that determines the clear boundaries of salient objects Rutishauser et al. (2004).

In computer vision society, the concept of visual saliency originates from the visual importance. The extracted saliency regions are valuable to assist various image understanding tasks, including, for example, object detection, content-based segmentation, image retargeting, and object recognition. However, without any priorknowledge, accurately isolating the salient objects from complicated environment still challenges the vision researchers.

In this work, we propose a novel saliency detection strategy, SMAP, which combines both the discriminant method and the stimulus-driven approach to emulate the human vision mechanism. In the proposed framework, the Gabor spectral residual is firstly introduced to locate the proto-background region. Based on the similarity measurement between the computing patch and background patches in the candidate pool, the local contrast is computed to generate the raw saliency map. We also incorporate the color distribution constraint to produce the full-resolution saliency map. The proposed algorithm outperforms the state-of-the-art methods even in the clutter environment where the background patches are full of texture information.

1.2 Object Recognition Via Deep Learning

Our visual world exists in a dedicated complexity. To understand scenes, the computers or other intelligent machines have to classify or recognize a nature image into different categories first. That is also the essential task for the human vision system. To realize the recognition, the rich attributes of visual entries should be uniquely encoded into reasonable representations. Although the visual scene is continuous, to precisely entitle the image into functional and semantic group remains a huge challenge in computer vision.

The main advances in object recognition were achieved thanks to the improvement in object representation learning. The performance of recognition schemes is heavily depended on the choice of features where the visual input applied. The manually engineered representations combined with discriminatively trained models have been among the best performing paradigms for related object recognition problems. However, such feature engineering is labor-intensive and most of the times, is not reliable to extract discriminative features for labeling the input.

In the recent years, the Restricted Boltzmann Machines (RBM) Hinton (2002) and Convolutional Neutral Network (CNN) LeCun et al. (1998) have emerged as powerful machine learning models. Adaptive Deconvolutional Network (ADN) Zeiler et al. (2011) is one of these edge-cutting deeply structured network. ADN is a multi-layer network which learns image representations that capture structure at all scales, from low-level edges to high-level object parts, in an unsupervised manner. Specifically, at each layer, the computing image/patch is decomposed into a linear combination of candidate features with sparse constraint. The inter-layer connection is in the form of max-pooling which responses the largest visual stimulus at a certain location. The original input image is always reconstructed at each layer. In this way, there is no information loss which exists in traditional Convolution Neutral Network, making the ADN more promising in hierarchical feature learning, and meanwhile benefiting the object recognition and categorization.

For ADN, despite the disentangling capability it has, we incorporate the $L_{1/2}$ norm regularization term instead of the original L_1 norm penalty to enhance the capability in feature learning. The proposed regularization forces the whole network to explore more sparse representations of the data and generate the hierarchical features with more discriminate information for object recognition.

1.3 Facial Landmark Detection via Low-rank Matrix Decomposition

To better understand the facial activity, we conduct facial feature parsing and landmark detection to assist the better analysis.

In computer vision, facial feature parsing refers to the task that segmenting face images into different facial feature components, e.g., eyes, nose and mouth, and applying related information analysis. The study of facial parsing is an attractive area due to its importance in multiple applications, including human identity recognition, animation, demographic analysis Guo et al. (2013), facial image synthesis Amberg et al. (2007) and face image sketching Wang and Tang (2009). All of these applications ask for accurate segmentation and more requirements to the parsing algorithm – robust to expression, pose and illumination variations. Most existing works accomplish the task by localizing landmarks on the input face as the initial points, and then refine them pixel-wisely by classification or regression till completely segment the regions of interest out. As the prior knowledge, the template matching model Liang et al. (2008) and graphic model Valstar et al. (2010) are applied to assist the parsing process.

In this work, we address the parsing problem from a new perspective and focus on facial feature detection from the face images instead of assigning label information for each pixel. Compared to previous methods, this detection-based approach is more efficient since it does not need to train the components descriptors piece-wisely. The facial features are treated as an entire set and can be detected at once. Specifically, our approach assumes a dataset of facial images with hand-labeled parsing map for each individual face. We emphasize that the alignment of all faces is not necessary. Clearly, the facial features contain discriminant shape, texture information, making them salient on the face region compared to the skin background. The intuitive idea to implement parsing is separating the salient components from the background. Our algorithm employs low-rank matrix decomposition method Liu et al. (2013) which

considers the skin background as the matrix spanning in low dimension subspace and the facial features with their discriminant characteristics performed as sparse noise. We also apply face detector to assist the face localization. In order to enhance the matrix decomposition, we introduce a transformation matrix T to force the algorithm learn the unique facial features.

1.4 Deep Tree-structured Face: A Unified Representation For Multi-task Facial Biometrics

Automatic facial image analysis has received considerable research interests due to its important role in computer vision and biometrics. As the key component, face feature is usually conducted under largely controlled environment and learnt for specific tasks which limit its discriminant capability in the unified representation. In this work, we present a novel deeply learnt tree-structured face representation to uniformly model the human face with different semantic meanings. The proposed feature is built from unsupervisedly learnt feature set, hierarchically combined region-by-region to generate a tree-structured representation. To enforce the semantic feature learning, we recursively apply semi-supervised AutoEncoder to incorporate label information which aims to disentangle the latent factors embedded in facial images. To validate the effectiveness of the proposed facial representation, we design comprehensive experiments based on FACES dataset which is considered as the most challenging one in terms of multi-factor overlapped. We show that the proposed feature yield unified representation in multi-task facial biometrics and the multi-task learning framework is applicable to many other computer vision tasks.

1.5 Contributions

The primary objective of the research is to provide a face centered image analysis system which is strengthened by several advanced technologies in computer vision. To approach this goal, the major contributions of this work can be summarized in the listed details:

- The novel three-level saliency based object detection method SMAP is proposed. Included in the methodology, the first level of SMAP accommodates statistical methods to generate proto-background patches, followed by the second level that implements local contrast computation based on image self-similarity characteristics. At last, the spatial color distribution constraint is considered to realize the saliency detection. The outcome of the algorithm is a full resolution image with highlighted saliency objects and well-defined edges. Quantitative evaluation based on a popular benchmark shows that the proposed approach has higher detection accuracy and more consistent performance for various categories of images;

- A revised Adaptive Deconvolutional Network (ADN) is studied as an approach to implement the object recognition. To strengthen the capability of the original deep network, $L_{1/2}$ norm regularization term is applied layer wisely to explore more discriminate features from images. Benefit from the new inference scheme of ADN, we visualize features learnt from each layer, and validate their roles in object recognition tasks. The hierarchical structure is evaluated based on the most popular benchmark dataset;

- It is the first time that low-rank matrix decomposition is introduced to solve the facial feature parsing problem. The proposed algorithm is detection-based method which is the initial work in this area. With the parsing results, we can easily extend the work to accomplish the facial landmark detection.

The high parsing accuracy guarantees the detection results receive competitive performance with the state-of-the-art;

- The unified deep face representation research is the first to propose the tree-structured face representation and implement it with designed semi-supervised AutoEncoder. It is proved to be effective in facial semantic learning. The proposed architecture is the first attempt to bridge the multi-task learning and deep learning to exploit latent feature learning for facial biometrics. It can be extended to many other computer vision applications.

1.6 Outlines

The organization of the dissertation is list as follows:

In chapter 2, the literature review is provided to introduce the state-of-the-art techniques in image analysis and facial biometrics included the salient object detection, object recognition deep learning neural network, so as the low-rank matrix decomposition theory. Chapter 3 explains the proposed salient object detection based on self-similarity. Chapter 4 introduces the $L_{1/2}$ norm regularized Adaptive Deconvolutional Network as a novel approach for object recognition. As a case study, facial expression recognition using ADN is studied in this part. Chapter 5 discusses the further facial activity analysis in terms of facial feature parsing and landmark detection. The further discussion about a unified face representation for multi-task facial biometrics is in Chapter 6. The entire work is concluded in Chapter 7.

Chapter 2

Literature Review

2.1 Review On Salient Region Detection

Finding and localizing an object or objects from 2-dimensional image is a fundamental task in computer vision. Human localize a multitude of objects in their vision field with little effort despite the position, type, color, contrast, size, perspectives and even the translation, rotation of the objects. Indeed, humans can distinguish between more than 30,000 visual categories, and can detect objects in the span of a few hundred milliseconds. However, if we want to transfer the ability from human to the vision machines, the detection task becomes crucial and challenged for many reasons. Successful algorithms and systems should adopt the large range of uncertainties included appearance changes, non-rigid transformations, scaling variations and object obstructions. In other words, the universal model does not exist for the generic detection problem.

One of the most common solutions for object detection/localization is to slide a window across the image, and classify each such local window as target or background locations. This approach has been successfully used to detect rigid objects such as faces and cars and has even been applied to articulated objects such as pedestrians. However, natural weakness of this algorithm exists in several aspects: the window

size which is determined by object scaling is a hyper-parameter and different from case by case. Without pre-knowledge about the detecting object, it is hardly to choose the window size and trial it by random pick; another problem is that, the classification operation which is involved to distinguish the windowed patch belonging is a supervised process, which means for one category of objects, we should train a specific model for it. It is not feasible to use the technique to detect multiple classes of objects. The representative researches belonged to this approach include Dalal and Triggs (2005), GIST Oliva and Torralba (2001) and Bag-of-Words Fei-Fei and Perona (2005) in object detection.

With the unsupervised preliminary, recent studies about object detection shift the focus to visual saliency. Visual saliency is the perceptual quality that makes a pixel, patch, object or person stand out to its neighborhood and thus attract human attention.

The study of the attention concept originates from human visual perception and neuro-psychology research. Researchers follow the methodology in Physiology to understand the eyes attention problem by analyzing the structure of human nervous system and brain. Although the mechanism to explain the operation of attention has not been completely understood, it shed light on computer vision groups that modeling the visual system could provide a rapid and reliable visual saliency detection.

The pioneer work about attention theory was conducted by William James (2013), where the key point proposed emphasized on the psychological response rather than the physical aspect. Following this direction, Broadbent Broadbent (2013) established the filtering theory of attention and Deutsch Deutsch and Deutsch (1963) proposed the vision response selection principles. In 1960s, Hubel and Wiesels famous work on cats vision research revealed the relationship between visual receptive fields and cortex Hubel and Wiesel (1962). At the same time, Treisman Treisman and Gormican (1988) proposed a theory which combines selection from early and late visual processes into a comprehensive model, referred to as the Feature Integrated Theory (FIT). The FIT model guided the biological attention research from theoretical reasoning

into computational implementation. In 1985, Koch and Ullman proposed so-called bottom-up saliency Koch and Ullman (1987), leading to the discovery of the underlying mechanisms of neutral vision system, where the bio-inspired features were used to highlight the saliency location. With the advanced technologies in biology, recent works about attention explored deeper in the V1 and V4 areas of the optical nerves Li (2002).

In par with the biological attention research, the other direction focuses on the study of computational saliency models, where a number of models have been constructed by adapting the FIT theory. Niebur and Koch were the first to realize the computational saliency map Niebur and Koch (1998) in 1996. Itti and his group refined Kochs work by generating a master saliency map considering various features such as color, intensity, orientation, etc. Itti et al. (1998). Some later models added more specific features, such as the symmetry pattern Kootstra et al. (2008), texture contrast Parkhurst et al. (2002), or motion information Itti et al. (2004) to the original structure. Saliency is also measurable in the frequency domain. In the spectral residual model Hou and Zhang (2007), saliency is described as the abnormal frequency from the smoothed FFT. This idea also incorporated natural image statistics related to the power law. In Achanta et al. (2009), the DOG filter was utilized to extract low-level features such as intensity and edges, which benefited the saliency evaluation. Notice that, most models mentioned above are stimulus-driven approaches where various features are crucial to determine the degree of saliency. The probability theory based on natural image statistics also gains popularity Itti and Baldi (2005); Zhang et al. (2008); Simoncelli and Olshausen (2001).

Among the aforementioned methods, the following algorithms are most state-of-the-art methods from different perspectives and mostly quoted in the peers works. The examination on properties of algorithms helps to reveal the advantages and limitations contained in each method.

In the approach of IT Itti et al. (1998), a 9 levels Gaussian pyramid is firstly created with successive Gaussian blurring and downsampling on the original input

image. Each feature is computed by a set of center-surround operations akin to visual receptive fields. Center-surround is implemented in the model as the difference between fine and coarse scales: The center is a pixel at scale $c \in 2, 3, 4,$, and the surround is the corresponding pixel at scale $s = c + d$, with $d \in 3, 4.$. The across-scale difference between two maps is obtained by interpolation to the finer scale and point-by-point subtraction. Totally three domain feature maps are calculated, which are colors, intensities and orientations of the image. At higher level, the calculated feature maps are fused together to generate the final saliency map with winner-take all strategy. After the hierarchical blurring and downsampling, the net information remained from original image contains few details and caused the saliency maps very blurred.

In the approach of MZ Ma and Zhang (2003), a low-resolution image is created by averaging the quantized blocks of the image. Each block is then represented by a single averaged pixel value which is a simulation of low-pass filtering. The resulted image is fed into the local contrast computation. The contrast value is computed by calculating the summation of the Euclidean distances between the current pixel and the surrounding pixels in LUV color space. After normalization, the saliency map is generated by the contrast values.

In the approach of SR Hou and Zhang (2007), the spectral residual of the image is computed by subtracting a smoothed version of the FFT log-magnitude spectrum from the original log-magnitude spectrum. The author advocates that the spectral residual represent the response towards the visual stimulus. By setting a hand crafted threshold, the direct component of the spectrum is filtered out. The remaining spectrum is applied inverse FFT to transfer into image space resulted in the saliency map.

In the approach of HC Cheng et al. (2015), the pixel saliency value is defined as a global histogram-based contrast. As an improvement over HC-maps, spatial relations is incorporated to produce region-based contrast (RC) maps where the input image is firstly segmented into regions, and then assign saliency values to them. The saliency

value of a region is now calculated using a global contrast score, measured by the regions contrast and spatial distances to other regions in the image.

Although extraction of salient objects by the aforementioned algorithms receives reasonable results in some aspects, the reliable and accurate saliency estimation remains challenging for computer vision society. Inherited from the advantages of the modeling of human visual system, the bio-inspired algorithms attribute a strong ability to precisely locate the attentive points related to the early stage responses of the visual neurons toward the stimulus in the field of receptive. However, despite of its precision, these points usually occupy only the blurred regions rather than the clear objects in the image domain, making subsequent applications inconvenient. Other computational strategies adopting multi-feature model are suffering to find out a general model that encompasses all diverse variations in saliency detection. Once the model failed to represent the salient feature, the computational result may generate unexpected saliency values. In other words, the model-based strategies cannot receive consistent performance considering the various application scenarios.

2.2 Deep Feature Learning Background

Recently, deep feature learning has been applied to a wide range of application scenarios Bengio et al. (2013). The most attractive attribute of the deep learning is the machines (RBM, CNN and AutoEncoder etc.) that learn a hierarchy of features from primitively low level to semantically high level, and significantly outperform existing approaches in areas like object recognition, music categorization, OCR and speech recognition tasks Bengio et al. (2013). As the structure of the network goes deeper, the learning machines are able to assemble the local features into a composition, increasing the tolerance to translation, rotation and scaling Zeiler and Fergus (2014). Meanwhile, with the standard pipeline, the deep structure is able to build invariance to capture domain knowledge, such as the facial morphology in case of facial expression recognition Rifai et al. (2012).

2.2.1 Restricted Boltzmann Machine

One of unsupervised deep learning network successfully applied in machine learning is using the Restricted Boltzmann Machine (RBM). The RBM is firstly proposed as a random neutral network based on statistical mechanics. It is an undirected bipartite network involved the Energy-based Model (EBM) Bengio (2009), and naturally develops from Boltzmann Machine (BM). We introduce the mathematics of EBM and BM here to understand the fundamental concepts of them.

Energy-based Model And Hidden Variables

The main objective of statistical modeling is helping to capture the dependencies between variables. Once these dependencies are determined, a model can be easily applied to inference the unknown variables given the value of the known variables. However, the barrier always exists in many cases that distribution of the observation data is not pre-acquirable. Energy-based model is helpful to solve it by associating a scalar energy to each configuration of the variables. The problem converts to modify that energy function so that it fits the desirable requirements Bengio (2009). The energy-based model defines the energy function as,

$$P(x) = \frac{e^{-Energy}}{Z} \tag{2.1}$$

where, the normalization factor Z is called the partition function. It defines as a sum running over the continuously input variable x,

$$Z = \sum_{x} e^{-Energy(x)} \tag{2.2}$$

and describes physical prosperities of the statistical ensemble.

In many of the application cases, we do not directly observe the variable x, and instead it will be reflected by some non-observable variables. Introducing the hidden variable enriches the power of the model LeCun et al. (2006). Considering the model

comprises an observation x and a hidden variable h, the energy-based probabilistic models define the new probability distribution as,

$$P(x,h) = \frac{e^{-Energy}}{Z} \tag{2.3}$$

Since x is observable, the marginal distribution is the main part we focus on

$$P(x) = \sum_{h} e^{-Energy(x,h)} \tag{2.4}$$

By introducing a new notation, free energy, the Eq. 2.4 can be mapped to the similar one as Eq. 2.1,

$$P(x) = \frac{e^{-FreeEnergy(x)}}{\sum_x e^{-FreeEnergy(x)}} \tag{2.5}$$

with $Z = \sum_x e^{-FreeEnergy(x)}$ and

$$FreeEnergy(x) = -\log \sum_{h} e^{-FreeEnergy(x)} \tag{2.6}$$

The data log-likelihood gradient then becomes easily to calculate. Using θ to represent the parameters of the model, and taking the gradient on the both sides of Eq. 2.5, we can obtain,

$$\begin{aligned} \frac{\partial \log P(x)}{\partial \theta} &= -\frac{\partial FreeEnergy(x)}{\partial \theta} + \frac{1}{Z}\sum_{\tilde{x}} e^{-FreeEnergy(\tilde{x})}\frac{\partial FreeEnergy(x)}{\partial \theta} \\ &= -\frac{\partial FreeEnergy(x)}{\partial \theta} + \sum_{\tilde{x}} P(\tilde{x})\frac{\partial FreeEnergy(\tilde{x})}{\partial \tilde{x}} \end{aligned} \tag{2.7}$$

The mathematical expectation of the Eq. 2.7 can be written as,

$$E_{\hat{P}}[\frac{\partial \log P(x)}{\partial \theta}] = -E_{\hat{P}}[\frac{\partial FreeEnergy(x)}{\partial \theta}] + E_P[\frac{\partial FreeEnergy(x)}{\partial \theta}] \tag{2.8}$$

On the right side of Eq. 2.8, the first term denotes the mathematical expectation taking over the training set, and the second term denotes expected value under the models distribution P. Therefore, to calculate the average log-likelihood gradient, we could sample from P and compute the free energy and then estimate with a Monte-Carlo way.

Boltzmann Machine

The Boltzmann Machine is a statistical model based on EBM and Restricted Boltzmann Machine is a particular form of Boltzmann Machine with more constraints on topological structure of the network. In the Boltzmann Machine, the energy function is formulated as,

$$Energy(x,h) = -b^T x - c^T h - h^T W x - x^T U x - h^T V h \tag{2.9}$$

The model parameters denoted by theta contain two parts: the biases b_i and c_i, and the weights W_{ij}, U_{ij} and V_{ij}. Following the tractable form of Eq. 2.7, the gradient of the log-likelihood can be written as,

$$\begin{aligned}
\frac{\partial \log P(x)}{\partial \theta} =& \frac{\partial \log \sum_h e^{-Energy(x,h)}}{\partial \theta} - \frac{\partial \log \sum_{x,h} e^{-Energy(x,h)}}{\partial \theta} \\
=& -\frac{1}{\sum_h e^{-Energy(x,h)}} \sum_h e^{-Energy(x,h)} \frac{\partial Energy(x,h)}{\partial \theta} \\
&+ \frac{1}{\sum_{x,h} e^{-Energy(x,h)}} \sum_{x,h} e^{-Energy(x,h)} \frac{\partial Energy(x,h)}{\partial \theta} \\
=& -\sum_h P(h|x) \frac{\partial Energy(x,h)}{\partial \theta} + \sum_{x,h} P(x,h) \frac{\partial Energy(x,h)}{\partial \theta}
\end{aligned} \tag{2.10}$$

Noted that $\frac{\partial Energy(x,h)}{\partial \theta}$ is easy to compute by taking derivative on Eq. 2.5. Therefore, the main calculation becomes to implement a procedure to sample from $P(h|x)$ and sample from $P(x,h)$. Then we can obtain an unbiases stochastic estimator of the log-likelihood gradient. The problem is solvable by constructing an Monte

Carlo Markov Chain (MCMC) Andrieu et al. (2003) or Gibbs sampling Geman and Geman (1984). Recent work solved it using even shorter chains as Contrastive Divergence Hinton (2002), and this method has been adopted in training of Restricted Boltzmann Machine.

2.2.2 Principles of Convolutional Neural Network

Convolutional Neutral Network (CNN) is one of the most attractive models in the recent development of cognition research Bengio (2009); Bengio et al. (2013); Bengio (2012). It was firstly inspired by the visual systems structure which is proposed by Hubel and Wiesel in their research of cats visual cortex Hubel and Wiesel (1962). Based on their works, the specific convolutional neural network efficiently reduces the complexity of back-propagation formed network. The CNN model based on the local connectivities between neurons and on hierarchically organized transformations of the image is efficient to obtain the translation invariant properties. Later on, researchers Alexander and Taylor improved the CNN theory and proposed the "improved perceptron which boosts the error-propagation algorithm Anthony and Bartlett (2009). LeCun followed-up with his idea to build up a new network structure trained with error gradient, obtaining the state-of-the-art performances in a variety of vision tasks LeCun et al. (1998, 2010). In this dissertation, an alternative method Decovolutional Network which improves the CNN is adopted to implement the object recognition Zeiler et al. (2011); Zeiler and Fergus (2014). It inherits the advantages from the CNN but empowers with the unsupervised learning method which makes it be more suitable in discovering discriminate features from image space.

In general, the basic element of CNN is composed by a two-layer structure. The first layer is called the convolutional layer. At this layer, the previous layers feature maps are convolved with learnable kernels and put through the activation function to form the output feature map. Each output map may combine convolutions with multiple input maps. Another layer is called sub-sampling layer. A sub-sampling

layer produces downsampled versions of the input maps. If there are N input maps, then there will be exactly N output maps, although the output maps will be smaller. Usually, the max-pooling is taken as the default down-sample scheme. On each element, the sigmoid function is selected as the activation function which is helpful to acquire the translation invariant LeCun et al. (2006).

The completed CNN is constructed by multiple network elements with a supervised perceptron as the output layer. The training method is following the traditional back-propagation algorithm to fine tune each parameter associated with every neuron.

2.2.3 Adaptive Deconvolutional Network

One particularly successful deep learning architecture is the Adaptive Deconvolutional Network (ADN) which offers appropriate features as well as the meaningful decomposition of the input images in multi-scaled levels Zeiler et al. (2011); Zeiler and Fergus (2014). The principle underlying of ADN is that, the features on each layer are learnt directly by minimizing the reconstruction error of the input image under a sparsity constraint from an over-complete set of feature maps. Unlike the traditional deep machines, which use the lower level reconstruction as the input for subsequent layers, there is no missing information resulted in error accumulation for ADN because of the layer-wised reconstruction for the original input. The unique setting 'switch' enforces the network to retrieve the max-pooling route, and thus makes the learning and inference reversible between input and its layered features. Considering its properties and advantages in feature extraction, we are the first to evaluate the effectiveness of the ADN based feature hierarchy in capturing the expression changes in facial images.

2.3 Low Rank Matrix Decomposition

To complete the face centered image analysis, we still need to parse and locate the landmarks on human face to get better face image understanding after we fully

awareness that there is a person exist in the image. In the work, Low-rank Matrix Decomposition is adopted to assist the landmark points detection. We detail the assumption, problem formulation and related techniques in the following paragraphs.

2.3.1 Mathematics of Low Rank Matrix Decomposition

Suppose that we have a matrix A of size $m \times n$ with rank-r, where $r \ll min(m, n)$. In many engineering problems, the entries of the matrix are often corrupted by errors or noise, some of them could even be missing, or only a set of measurements of the matrix can be accessible instead of the matrix entries directly. In general, we model the observed matrix D to be a set of linear measurements on the matrix A, subject to noise and gross corruptions i.e., $D = L(A) + \eta$, where L is a linear operator, and η represents the matrix of corruptions. The problem is seeking to recover the genuine matrix A from D.

When to consider the case where L is the identity operator and η is a sparse matrix but whose non-zero entries can be practically unbounded. Since the rank r of A is unknown, the problem is to find the matrix of lowest rank that could have generated D when added to an unknown sparse matrix η. Mathematically, for an appropriate choice of parameter $\gamma > 0$, we have the following combinatorial optimization problem to solve,

$$\min_{X,E} rank(X) + \gamma \|E\|_0 \qquad subject \quad to \quad D = X + E \tag{2.11}$$

where $\|\|_0$ is the L_0 norm.

Since the solving the above problem is NP-hard, one can alternatively solve it by another convex surrogate,

$$\min_{X,E} \|X\|_* + \lambda \|E\|_1 \qquad subject \quad to \quad D = X + E \tag{2.12}$$

where $\|\|\|_*$ represents the matrix nuclear norm and is the best convex approximation to the rank function. $\|\|\|_1$ represents L_1 norm, and λ is a positive constant. In the paper proposed by Wright Wright et al. (2009), it shows that X and E can be perfectly recovered from Eq. 2.12

Chapter 3

Saliency-based Object Detection

3.1 Saliency Region and Visual Saliency Analysis

The hypothesis of human attention theory Li (2002); Reynolds and Desimone (2003) points out that the human visual system selectively analyzes the details of the partial image in the vision field but ignores the majority rest. From the general statistical analysis of natural images, we also found that, only a small portion of the single image contains richer information. The straightforward assumption is thus that the interesting part in vision has a corresponding relationship with the image region that has more information. All the other patches sharing similar appearance are treated as redundant Zhang et al. (2008). The similar idea appeared in Gao's work Gao et al. (2008). However, in his work, calculating the pre-defined clusters through a supervised method is neither efficient nor affordable with limited database or computational resource.

Take the image in Fig. 3.1 as an example where we manually select the patches from the foreground and background objects and calculate the mutual information between these patches, as shown in Table 3.1. We observe that the foreground patch (F in Fig. 3.1) contains more distinct information than the background patches (R1,R2,Y1,Y2 in Fig. 3.1).

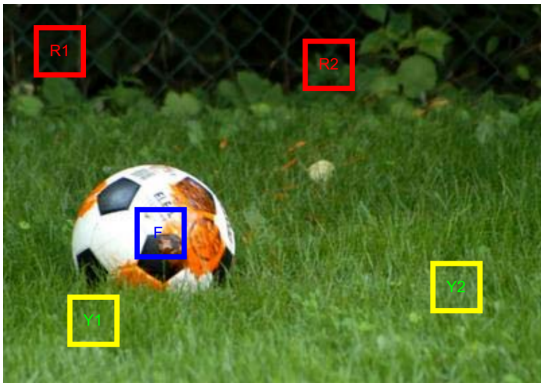


Figure 3.1: The background patches have the self-similarity attribute

Table 3.1: Mutual information (MI) between the labeled patches in Fig. 3.1. All patches from the background share similar appearance with MI less than 3.704; the MI between the foreground and background patches are more than 5.25.

M I	F	R1	R2	Y1	Y2
F	0	5.2585	6.2765	6.6777	6.8207
R1	5.2585	0	2.3687	2.5628	2.6069
R2	6.2765	2.3687	0	3.4273	3.3272
Y1	6.6777	2.5628	3.4273	0	3.7082
Y2	6.8207	2.6069	3.3272	3.7082	0

From the perspective of coding theory, one can always decompose the information of a static image into two parts, the prior knowledge and the abnormal properties Hou and Zhang (2007). The former, most of time, is redundant and supposed to be suppressed by the coding process. The latter normally carries more distinctive information and therefore is the main focus of our research. There have been quite some efforts devoted to the search of the 'disctinctive' information in the image from different aspects. Until now, there has been no general model proposed that comprehensively describes the varieties of the whole image.

In this paper, we take the inverse approach to conventional saliency detection by first detecting the parts in the image that are *not* attractive. This approach would identify patches that do not contain distinctive attributes and share the self-similarity across the image, and thus are deemed as background candidates, referred to as the "proto-background", as opposed to "proto-foreground" detected in conventional

approaches. Because of the self-similarity attributes, the accuracy of background detection is better appreciable than the detection of foreground salient parts.

Following the early selection of perception which outputs only the attentive points directly responding to outside stimulus, the subsequent biological process of human vision is the 'refinement process' that generates the perceptive field containing the semantic objects. To generate a perceptive field containing the salient objects with clear boundaries, local contrast is incorporated to evaluate the distinctiveness of each pixel. We define the local contrast as a feature-based similarity function between the evaluated patch and the proto-background patches.

Another plausible feature to describe saliency is the color distribution. It is commonly accepted that the color spatial distribution should be concentrated rather than scattered in the salient object that are attractive. As a global feature, the color distribution constraint also assists saliency representation to achieve a uniform and consistent performance among all images.

3.2 The Saliency Detection Methodology - SMAP

The goal of the proposed saliency detection system is to locate the potentially interesting foreground and to emulate the refinement process of eyes to better represent the saliency region and extract salient objects with full resolution. To achieve this goal, the pre-attentive process should filter the proto-background out of the image. Next, local contrast calculation is conducted to generate the raw saliency map. By adapting the observation that the color distribution of saliency object cannot be widely spread, we introduce the color distribution as a global constraint to assist the final detection. Correspondingly, the framework of the proposed system is composed of three parts and the saliency map produced would benefit in accuracy and uniform performance from both the local contrast and global constraint calculation.

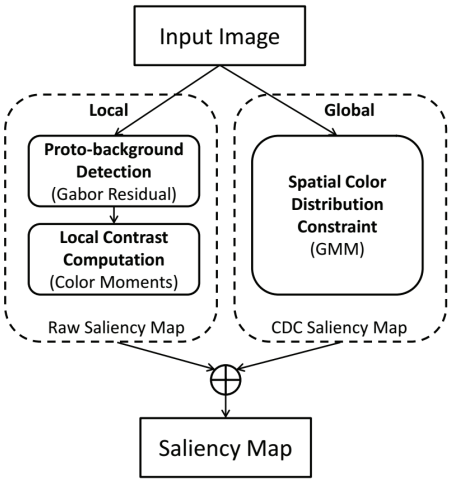


Figure 3.2: The diagram of the proposed saliency detection system

3.2.1 The Local Stimuli Response: Proto-background Detection

Image analysis by Gabor filter bank is considered to resemble the perception in the human visual system Daugman (1985), where the quantitative response and tuning mechanism along the ventral stream of visual cortex is well modeled by the Gabor wavelets Riesenhuber and Poggio (1999). We thus adopt a set of Gabor filters at different scales and orientations to obtain the early attentive response. The sinusoidal Gaussian property enables the filters to produce the scale and position-tolerant features. The explicit form of the 2-dimensional Gabor filter in the spatial domain is described by,

$$G(x,y) = \exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) \cos(\frac{2\pi}{\lambda} x') \tag{3.1}$$

where $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$ are the rotation factors of the Gabor filter controlled by the angle θ. σ is the standard deviation of the Gaussian envelope and γ is the spatial aspect ratio, which is fixed to $\gamma = 0.3$. λ represents the wavelength of the sinusoidal factor. Tuning of λ relates to the change of the functional scale of the Gabor filter. Features in six orientations at five scales are computed. This

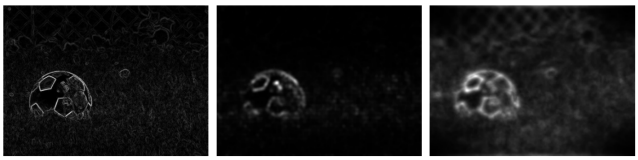


Figure 3.3: Gradient, spectral residual Hou and Zhang (2007) and Gabor residual of the image from Fig. 3.1.

filter bank is designed to acquire strong responses at locations where sharp stimulus matches at different orientations Guo et al. (2009).

In some natural scene images, the out-of-focus effect often causes blur at the object boundaries, introducing more uncertainties to the detection. To be more tolerant against the shift and out-of-focus effect, we measure the spectral residual from the Gabor spectrum. In one dimension, the spectral residual is calculated as,

$$R(f) = \ln |G(f)| - h_n * \ln |G(f)| \tag{3.2}$$

where $h_n = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$ works as an average filter, $G(f)$ is the real part of the Gabor spectrum of the input image. The $iFFT$ operation applied on $R(f)$ would create an image with highlighted regions that relate to the early attentive points known as 'abnormal', or more specifically, the 'proto-object'. Because the response is only analogous to the lowest level of early attention in the human visual system, it carries little semantic information but the maximum contrast caused by stimulus. Different from the classic spectral residual approach Hou and Zhang (2007), Gabor residual enriches the response in multiple orientations and scales. The texture details captured in the Gabor spectral space directly relate to the primitive receptive fields of the human vision. Fig. 3.3 indicates the relationship between the gradient and the Gabor residual. Clearly, Gabor residual has stronger response towards the gradient changes than conventional spectral residual.

Based on the proto-object calculated, the proto-background detection process is shown in Fig. 3.5. All the input images are uniformly rescaled to the size of 256 × 256. Then the Gabor spectral residual algorithm is applied. We divide the resulting

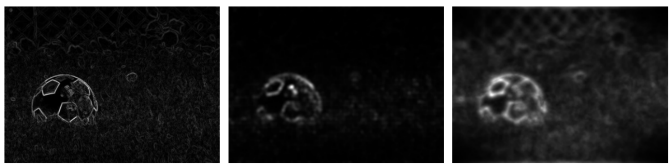


Figure 3.4: Various division thresholds and the resulting Gabor residual images. From left, the division threshold is set to 0.2 to 0.8 with 0.2 as the interval. 0.6 is the default setting.

mask into quantized blocks of size 32×32. Counter-intuitively, the blocks with average intensity *below* a pre-defined threshold are selected as the proto-background regions. The selected blocks are stored in the Background Candidate Pool (BCP). The advantages of choosing the proto-background instead of the proto-object are two-fold. On one hand, the 'redundancy' property in the proto-background is more generic in different images. On the other hand, the scheme is robust to inaccurate detection. To balance the accuracy between the proto-objects and proto-background detection, we set the threshold as 0.6 times the average pixel intensity of the Gabor spectral residual image to select proto-background patches from the Gabor residual images. Higher threshold produces more precise proto-objects but poor proto-background estimation.

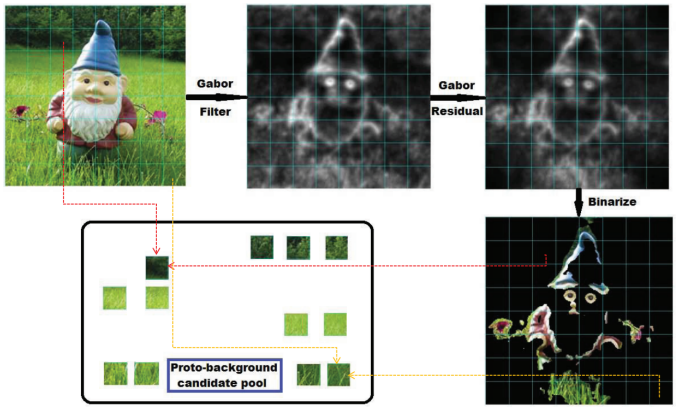


Figure 3.5: The process of generating the background candidate pool

3.2.2 The Fine Tuning Process: Local Contrast Calculation

After locating the proto-background, the refining process is conducted to determine the whole area of the salient targets.

Almost all saliency algorithms utilize the color channels in different color spaces. The *RGB* color decomposition is the most frequently employed. Others argue *Lab* provides better approximation as its components more closely match the human perception in lightness and chromatics Borji and Itti (2012). Here we adopt the *HSV* color space of Hue, Saturation and Value since it accommodates more traditional and intuitive color mixing models based upon how colors are organized and conceptualized in human vision Myers (1979). One favorable advantage received by using the HSV color decomposition is that, the saliency value calculated does not rely on any specific color.

Motivated by the color indexing technique, we incorporate the color moments to differentiate image patches based on their color feature. The color distribution of an image can be interpreted as the probability distribution. Thus, the moments are always proper choice to represent this distribution. We propose the first three moments *mean*, *standard deviation* and *skewness* as the image color index. If the pixel value of a given color distribution is defined as p_i in HSV color space, the moment metrics can be defined as,

$$\mu = \begin{pmatrix} \mu^H \\ \mu^S \\ \mu^V \end{pmatrix}, \quad \sigma = \begin{pmatrix} (\frac{1}{N}\sum_{i=1}^{N}(p_i^H - \mu^H)^2)^{\frac{1}{2}} \\ (\frac{1}{N}\sum_{i=1}^{N}(p_i^S - \mu^S)^2)^{\frac{1}{2}} \\ (\frac{1}{N}\sum_{i=1}^{N}(p_i^V - \mu^V)^2)^{\frac{1}{2}} \end{pmatrix}$$
$$s = \begin{pmatrix} (\frac{1}{N}\sum_{i=1}^{N}(p_i^H - \mu^H)^3)^{\frac{1}{3}} \\ (\frac{1}{N}\sum_{i=1}^{N}(p_i^S - \mu^S)^3)^{\frac{1}{3}} \\ (\frac{1}{N}\sum_{i=1}^{N}(p_i^V - \mu^V)^3)^{\frac{1}{3}} \end{pmatrix} \tag{3.3}$$

where μ is the mean value of the distribution, N is the total amount of pixels in the distribution and the superscript represents different color channels. The three moments physically evaluate the *average*, *variance* and *degree of asymmetry* in color distribution. An image (or a patch I) is then easily characterized by totally 9 moments in the 3 color channels, i.e., $I = (\mu, \sigma, s)^T$.

The similarity measurement is defined as a sum function of the weighted difference between the moments of two distributions, H and I, i.e.,

$$d_{sim}(H, I) = \omega_1^T \cdot \Delta\mu + \omega_2^T \cdot \Delta\sigma + \omega_3^T \cdot \Delta s \tag{3.4}$$

where $\Delta\mu$, $\Delta\sigma$ and Δs represent the difference of moments between two distributions. Notice that the similarity comparison happens within the single image. The environmental condition is supposed to be unchanged. The weight vector is set to $\{\omega_i = [1, 1, 1]\}_{i=1,2,3}$, with nondiscriminant treatment on every element. Although these statistical representations vary significantly, they all help capture the color, edge features, repetitive patterns and complicated texture in a unified way, where the self-similarity is considered.

After acquiring the proto-background blocks, we subdivide these blocks into smaller 7×7 cells with 50% overlap. When evaluating the local contrast of a pixel p_i, the moments of 3×3 patch centered at p_i are computed. The similarity between p_i and the background cells are calculated according to Eq. (3.4). We can obtain a series of similarity values. The local contrast of p_i is defined as the accumulated minimal 128 similarity values between the computing patch H_i and the patches in BCP. H_i is the 8-neighbor patch of pixel p_i. The image texture with similar color property and repetitive patterns is easily matched with the moment vectors in BCP, and thus receives a low contrast value. Apparently unique patches become visually salient since its similarity measurement based on BCP is rather strong. The image texture with similar color property and repetitive patterns is easily matched with the moment vectors in BCP, and thus receives a low contrast value. Apparently unique

Figure 3.6: Raw saliency map demonstration. From top row to the bottom: original images, raw saliency maps.

patches become visually salient since its similarity measurement based on BCP is rather strong.

The contrast values are normalized into the range [0 1]. The local contrast of the input image is treated as the raw saliency map S_{raw}. The figures in the $2nd$ row of Fig. 3.6 illustrate the computed raw saliency maps S_{raw}.

3.2.3 The Global Saliency Response: Color Distribution Constraint

Although the Gabor spectral residual is quite effective to locate the saliency points in the receptive fields, the frequency domain method still suffers from one deficiency since the spatial distribution information is ignored. One of the possible drawbacks is that, in the clutter environment, the detected saliency parts may scatter all over the image.

The spatial distribution of a specific color may contribute significant information to the saliency detection. The concentration property of the saliency object indicates that, the wider a color distributes, the less possible it attracts human vision.

It is assumed that the color of the saliency object concentrates around a small region, making the small variance more attractive. Thus, the color distribution constraint indeed provides another important feature for saliency.

The Gaussian Mixture Model (GMM) is introduced to represent colors Liu et al. (2011); Schwarz et al. (1978); Calinon et al. (2007). The image is treated as a data set with each pixel being a data point. A mixture model of c color clusters indexed by j, $j = 1, 2, ...c$, is defined by a probability density function,

$$P(p_i) = \sum_c P(j)P(p_i|j) \tag{3.5}$$

where p_i is a vector represents the pixel value, $P(j)$ is the prior and $P(p_i|j)$ denotes the conditional probability density function. The image is modeled by a mixture of c Gaussians of dimension d (for color images, $d = 3$). The parameters in Eq. (3.5) become

$$\begin{aligned} P(j) =& \pi_j \\ P(p_i|j) =& \Phi(p_i|\mu_j, \Sigma_j) \\ =& \frac{1}{\sqrt{(2\pi)^d|\sum_j|}} \exp\{-\frac{1}{2}(p_i - \mu_j)^T\Sigma_j^{-1}(p_i - \mu_j)\} \end{aligned} \tag{3.6}$$

where $\{\pi_j, \mu_j, \Sigma_j\}$ represent the prior, mean and covariance matrix of the color cluster j in GMM respectively. The probability of a pixel p_i assigned to the color cluster j is defined as,

$$P(j|p_i) = \frac{P(j)\Phi(p_i|\mu_j, \Sigma_j)}{P(p_i)} \tag{3.7}$$

The standard Expectation-Maximization (EM) algorithm is applied to solve the parameter estimation for mixed Gaussians iteratively. However, an obvious shortcoming of EM is that the optimal number of clusters in one image is unknown beforehand. To encode the image dataset with any fixed number of clusters will

result in either the deficiency in data modeling or parameter over-fitting. A traditional strategy to trade-off between optimizing data's likelihood and minimizing the number of parameters used is the model selection. Bayesian Information Criterion (BIC) Schwarz et al. (1978) is then incorporated to tackle the estimation problem. The BIC score provides selection criteria to determine the optimal number of GMM clusters with the definition,

$$S_{BIC} = -2\mathcal{L} + n_p \log(N) \tag{3.8}$$

where $\mathcal{L} = \sum_{i=1}^{N} \log(P(p_i))$ and n_p is the number of free parameters in the model. For GMM, $n_p = (j-1) + j(d + \frac{1}{2}d(d+1))$. N is the total number of data points. $\mathcal{L}$ is the log-likelihood function which measures the fitness of modeling on the data. $n_p \log(N)$ works as a penalty term to control the complexity of the model. Considering the goal of the algorithm is to segment the image based on color, we select the optimal number of color clusters as,

$$K = \begin{cases} \underset{c}{argmin} S_{BIC} & \text{if } c <= 6 \\ \lceil (\underset{c}{argmin} S_{BIC})/2 \rceil & \text{if } c > 6 \end{cases} \tag{3.9}$$

where operator '$\lceil \rceil$' represents the smallest integer greater than x. The selection of K reaches a tradeoff between the optimal data fitting and reduced modeling complexity. We compute with a set of candidate Gaussians up to 15 color clusters and select the model parameters according to Eq. 3.9.

The spatial distribution variance for each color cluster can then be interpreted as the weighted offset between each pixel position and the centroid. We penalize the offset if it is larger than half of the image range. The horizontal variance is formulated as

$$V_h(c) = \frac{1}{|X|_c} \sum_i P(c|p_i) \frac{\lambda \cdot |x_h - m_h|^2}{1 + \exp(-\gamma \cdot width \cdot |x_h - m_h|^2)} \tag{3.10}$$

where m_h is the x-coordinate of the centroid for color cluster c, x_h is the x-coordinate of pixel p_i, λ and γ are two positive constants which control the scope of the sigmoid-like penalty function, $width$ denotes the width of the input image, $|X|_c$ is the normalization term which is the total intensity value of color cluster c. With similar notation, the vertical color distribution variance $V_v(c)$ should be computed in the same way. In the following experiments, we use $\lambda = 100$ and $\gamma = 0.5$ as the default settings for this paper. Once pixels within the color cluster scatter across the image more than half of the image range, then the strength of the penalty would increase gradually.

Associated with the color distribution variance, the color distribution-constraint (CDC) feature map is defined as

$$S_{cdc}(p_i) = \sum_c P(c|p_i)(1 - V_h(c))(1 - V_v(c)) \tag{3.11}$$

In Fig. 3.7, we evaluate the penalty effect using a toy example. The examined image contains four color clusters. Except for the background color, the two squared blocks and a horizontal bar exist within the image domain with three distinctive colors. Without the spatial color distribution constraint, the squared blocks and the bar should have the same saliency value since they have the same area of coverage. Considering the concentration attribute of the salient object, the sigmoid-like term penalizes the bar since the horizontal variance of the bar is larger than half of the width of the image. From the demonstrated results, it is clear that the squared blocks are more salient than the crossing bar after applying the color distribution constraint in Eq. 3.11.

We normalize the feature map to the range of [0 1] which is comparable to the raw saliency map S_{raw}. The color distribution constraint saliency maps S_{cdc} are demonstrated in the $2nd$ row of Fig. 3.8.

Figure 3.7: A toy example to demonstrate the penalty effect. From left (a) original image, (b) the CDC saliency map without spatial penalty, (c) the CDC saliency map with sigmoid-like penalty term.

Figure 3.8: Color distribution-constraint saliency map demonstration. Top row: original images; bottom row: color distribution-constraint saliency maps.

3.2.4 Saliency Map Generation

The last process to generate the saliency map involves a linear combination of the raw saliency map S_{raw} generated by local contrast and the color distribution constraint map S_{cdc} to produce the final saliency map S_{map}.

$$S_{map}^{\alpha} = \alpha \cdot S_{raw} + (1 - \alpha) \cdot S_{cdc}(I) \quad \alpha \in (0,\ 1) \tag{3.12}$$

where α is a user controlled parameter which tunes the contributions of the two factors. Empirical study on α reveals that larger α ($\alpha > 0.5$) performs well on natural outdoor scene. However, without any presumption about the database, various α values may result in unstable performance of the model. To optimize the bias between local contrast and color distribution, and minimize the chaotic information in the saliency detection, we propose the 2-dimensional entropy as the criterion to evaluate the linear combination of the two maps. Unlike the conventional entropy only considering the probability of information, 2D entropy incorporates spatial information which represents the object geometric characteristics contained in the image. The optimal α is determined by

$$\alpha_{opt} = \underset{\alpha}{argmin}\{\mathcal{H}(S_{map}^{\alpha})\} \quad \alpha \in (0,\ 1) \tag{3.13}$$

where $\mathcal{H} = -\sum_i \sum_j p_{ij} \log p_{ij}$, and p_{ij} represents element in 2D histogram of the image S_{map}^{α}. Smaller entropy value illustrates that the texture/edge information is largely suppressed and salient pixels are concentrated into small regions.

3.3 Experiments and Evaluation

We evaluate the performance of the proposed saliency model SMAP in two ways: 1) predicting human visual fixations, and 2) detecting the saliency objects which humans pay attention to. For these purposes, two standard databases, the MIT database and the MSRA database, and the state-of-the-art saliency detection techniques are

selected for comprehensive comparison. The MIT database collects eye-tracking data of 15 views who free-viewed 1003 natural indoor and outdoor images Judd et al. (2009). The MSRA database (subset) contains 1000 images from various categories such as plants, animals, traffic signs, human sports, etc. It also provides binary masks of human labeled saliency ground truth. We select 9 state-of-the-art saliency detection algorithms for comparison purpose. These algorithms cover the recent popular techniques of global (HC Cheng et al. (2015), LC Zhai and Shah (2006)), local (IT Itti et al. (1998), MZ Ma and Zhang (2003)), frequency domain (SR Hou and Zhang (2007), AC Achanta et al. (2008), FT Achanta et al. (2009)), multi-modality (GB Harel et al. (2006), CA Goferman et al. (2012), RC Cheng et al. (2015)) and learning-based (LR Shen and Wu (2012)) approaches.

3.3.1 Human Visual Fixations Prediction

Human visual fixations, also known as attention points, directly relate to the primitive level response towards the stimulus. Facilitated with the Gabor spectral residual in local contrast calculation, saliency maps generated by SMAP should be effective to cover the human fixations. We evaluate SMAP by employing MIT's database and compare it with the original work in Judd et al. (2009).

Unlike the proposed SMAP, the output of the MIT work is the gray level image containing the saliency region rather than the clear-bounded salient objects. To perform a fair comparison with the proposed SMAP, we design the experiment by comparing the 100 most salient pixels with the fixation points. If the fixation point meets the salient pixel at the same location, we count it as a hit. The percentage of total number of hits over the number of fixations is used as the evaluation metric.

The MIT database has a strong bias on "centering" which means the salient objects most likely locate at the center of the images. Considering this prior knowledge, paper Judd et al. (2009) applied a convolution with Gaussian kernel on the saliency map to compensate for the centering effect. Another solution to offset

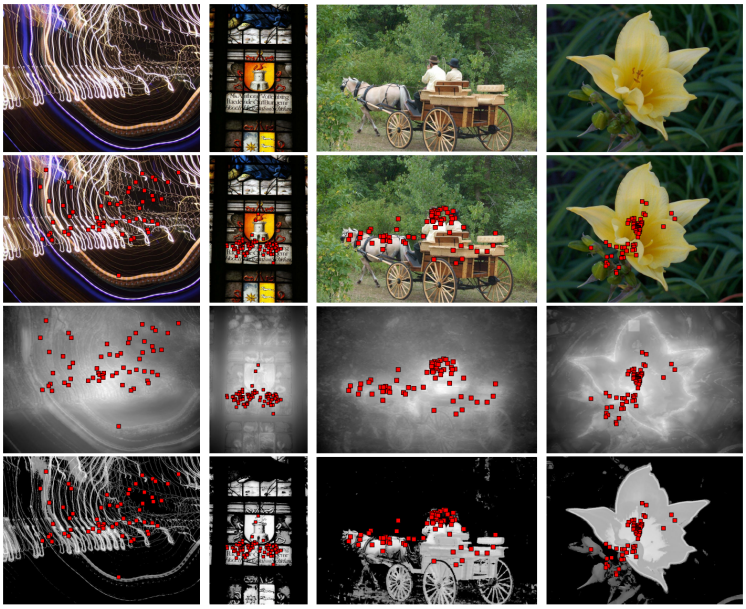


Figure 3.9: Human visual fixation comparison. From top row, the original images, images with human fixation points (red dots), saliency maps from Judd et al. (2009) with fixations and SMAP with fixations.

the centering effect was reported in Judd et al. (2012), where they linearly combined the saliency map with a Gaussian map. In this paper, we adopt both methods in SMAP with the same scheme but slightly change the weights in linear combination. We reduce the weight on centering map to $w = 0.2$. The quantitative comparison is shown in Fig. 3.10. Clearly, the top 10% salient pixels from the SMAP linearly combined with the centering map cover nearly 90% of the fixation points. Compared with the diffused maps given by Judd et al. (2009), SMAP also outperforms in term of the clear boundaries and complete salient objects detected.

3.3.2 Visual Saliency Evaluation with Extracted Attention View

Unlike the attention fixation points Ma and Zhang (2003) which are analogous to the primitive level of human attention caused by visual stimulus, attention view attracts more research interests by emphasizing on the detection of the whole objects

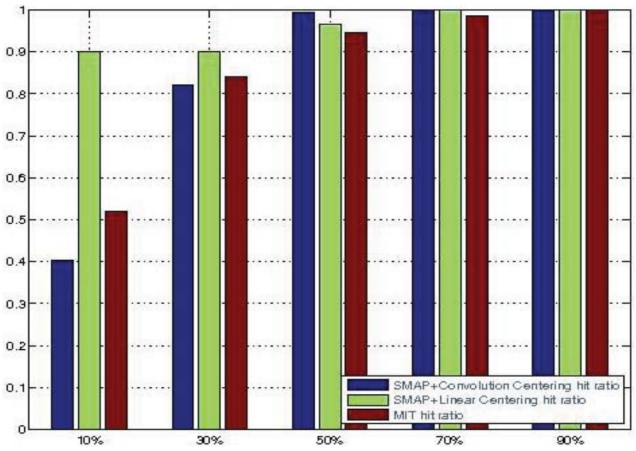


Figure 3.10: Quantitative comparison for human visual fixation prediction. Using hit ratio as the measurement metric.

rather than several isolated points from image. To generate rectangles containing the complete saliency objects, we perform a raster scanning on the binarized SMAPs. The output rectangles should contain at least 95% salient pixels in SMAP Liu et al. (2011). If the image containing more than one center of attraction, the searching algorithm automatically repeats and generates several rectangles for different objects in the image.

The evaluation of the extracted attention view is carried out by user experiments Ma and Zhang (2003). We set 3 assessment levels, GOOD, ACCEPTABLE and FAILED, for the users to evaluate the extraction results, since quantitative measurement is difficult to assign to these assessments. Generally, the GOOD cases include precise detection with 80% accuracy. The ACCEPTABLE cases should cover 50% of the saliency objects, otherwise, it will be labeled as FAILED.

Human visual saliency prediction contains its own consistency. Therefore, three users are involved to evaluate the extraction of attention view by feeding the images with rectangles on the MSRA dataset. According to Judd et al. (2009), three participants help to maintain the saliency detection accuracy to reach nearly 90%. The statistical results are displayed in Table 3.2.

Table 3.2: The quantitative evaluation of user experiment on extracted attention view.

Users.	GOOD	ACCEPTABLE	FAILED
1	88.81%	5.97%	5.22%
2	87.31%	6.72%	5.97%
3	86.57%	8.21%	5.22%
Avg.	**87.56%**	6.97%	5.47%

Figure 3.11: Visual saliency detection results. The red rectangles are extracted attention view areas calculated based on SMAP. The yellow rectangles are ground truth areas calculated based on ground truth mask with exhaustive search algorithm.

The average ratio of GOOD assessment reaches as high as 87.56%. The subjective experiments showed the effectiveness in human attention view extraction strategy. The visual saliency detection results are demonstrated in Fig. 3.11.

3.3.3 Visual Comparison on Different Types of Images

The first experiment in this section is designed to compare the color independency property as the saliency at a visual location should be irrespective of the actual color feature Li (2002). We choose the images which contain both multi-color saliency objects and uni-color objects located in multiple color background as illustrated in Fig. 3.14(a)(b)(c). Beside the Gabor spectral residual helping capture the color changes, the local contrast scheme makes the repeat texture and color patterns

redundant and thus decreases the degree of the saliency. The comparison is obvious that the proposed SMAP method is color independent.

The SMAP algorithm also presents reliable performance to extract the saliency from clutter environment, where the background of the images are full of texture information. Most frequency domain methods would fail in detecting saliency objects due to the high frequency texture in the background. The global histogram based algorithms are also influenced a lot by the widely distributed texture. For this type of images, the self-similarity property helps characterize the redundant texture. This in turn facilitates the removal of the redundancy easily by the statistical attributes which is calculated according the proto-background candidate pool. In Fig. 3.14(d)(e), we demonstrate the effectiveness of the proposed algorithm in complicated environment.

In some extreme cases, the saliency foreground shares a similar appearance with the background. That is, the contrast degrades and can be easily ignored. The experimental results in (Fig. 3.14(f)(g)) further demonstrate that our method holds an advantage to explore the subtle low contrast information than the peer.

3.3.4 Quantitative Comparison on Image Segmentation Results

To obtain a quantitative evaluation of the proposed method, we compute the binary maps using thresholds ranging from 1 to 250 with the interval of 10 on the SMAP. The precision and recall Achanta et al. (2009); Cheng et al. (2015) are calculated and compared with on the whole benchmark database. This segmentation enhancement test reflects the overall effectiveness of the SMAP algorithm. Fig. 3.12 shows that SMAP outperforms the others.

Another objective comparison uses the adaptive threshold T_a as in Eq. (3.14) instead of the fixed ones to binarize all the saliency maps, where $S(x, y)$ denotes the pixel value at (x, y) of the saliency map, $Width$ and $Height$ are the dimensions of

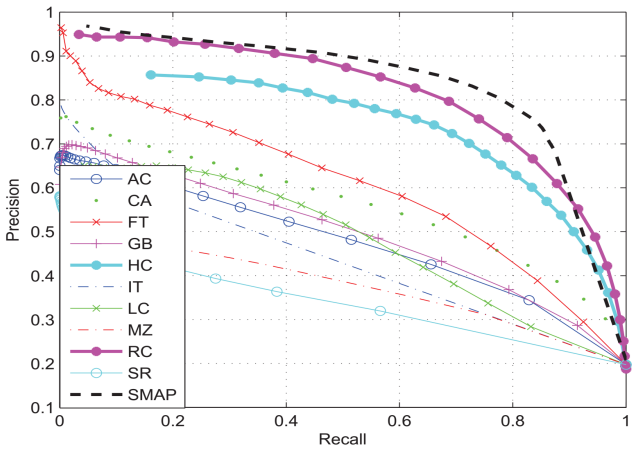


Figure 3.12: Precision and Recall curve comparison with the state-of-the-art algorithms. SMAP is the proposed algorithm.

the image. Average precision, recall and F-Measure are evaluated over the entire database.

$$T_a = \frac{2}{Width \times Height} \sum_{x=0}^{Width-1} \sum_{y=0}^{Height-1} S(x,y) \tag{3.14}$$

$$F_\beta = \frac{(1+\beta^2) Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{3.15}$$

We use the setting $\beta = 2$ to emphasize on recall rather than precision. The comparison results are shown in Fig. 3.13. Notice that, the RC combines graph-cut segmentation method and thus narrows the contrast to a region-based scheme. With this geometric prior, it receives relatively higher precision value.

The previous experiments demonstrate the effectiveness of the SMAP method. We further investigate in the roles of the local raw saliency map (RAW) and the global color distribution-constraint map (CDC) in terms of the F_2 value to compare with other algorithms. Clearly, the RAW and CDC maps contributes both the precision and recall which guarantee the saliency maps generated by the SMAP algorithm receive the highest F value. The highest recall ratio also demonstrates the SMAP detection best covers the whole saliency region.

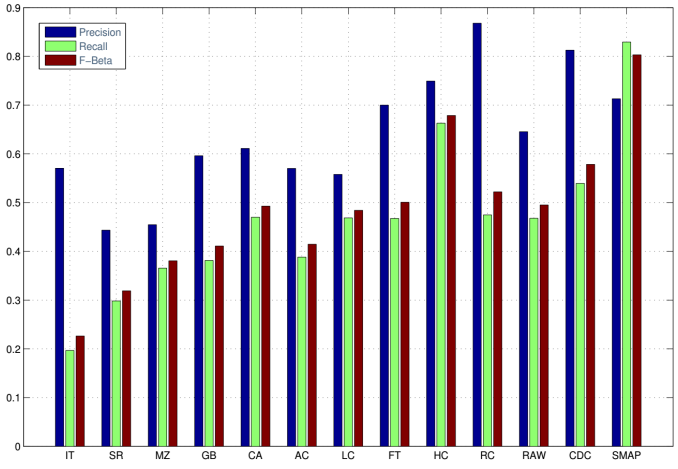


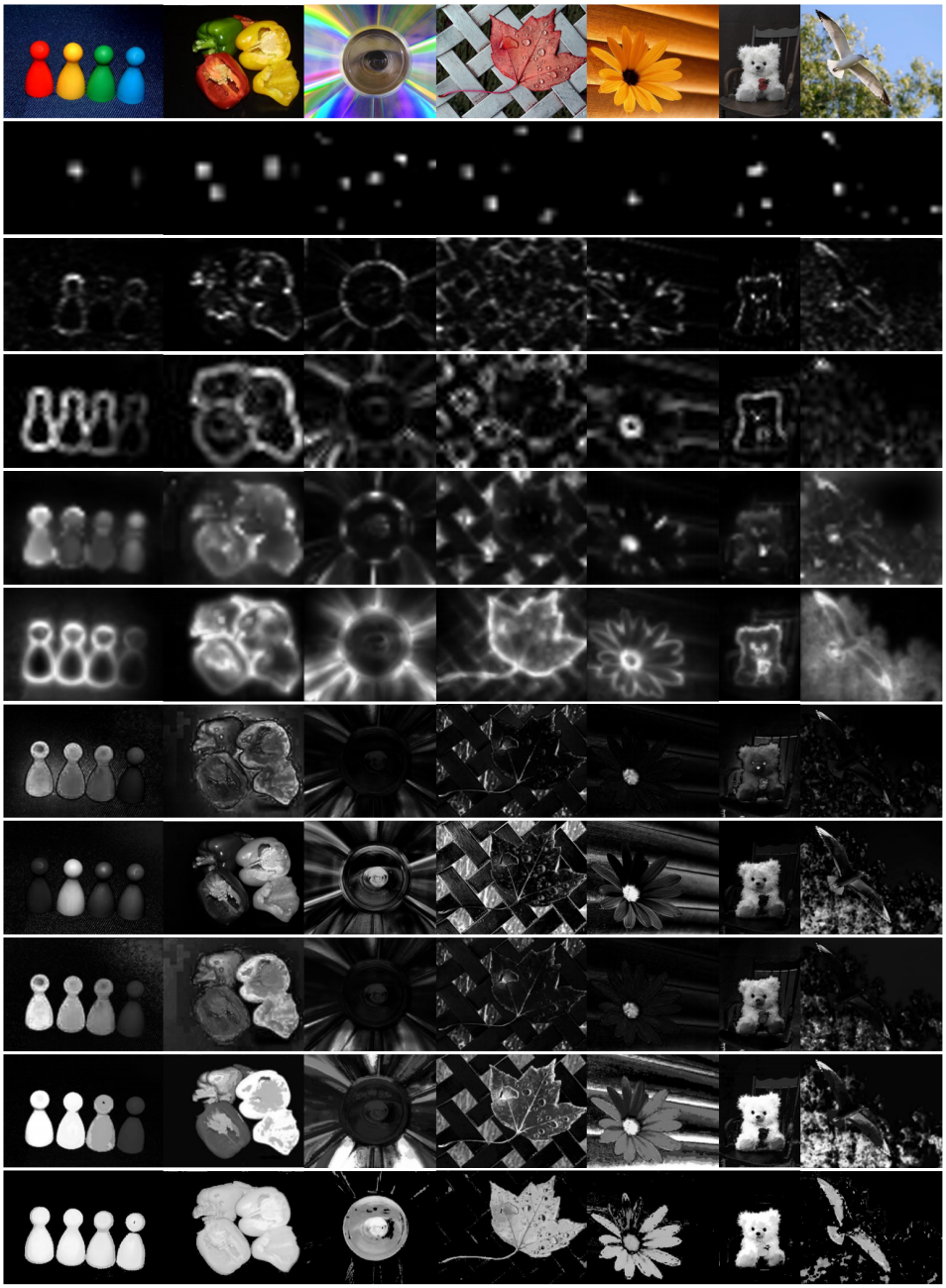
Figure 3.13: F-measure evaluation.

3.4 Applications

Precise saliency detection benefits many computer vision applications. In this section, three interesting applications that the saliency map can help are discussed.

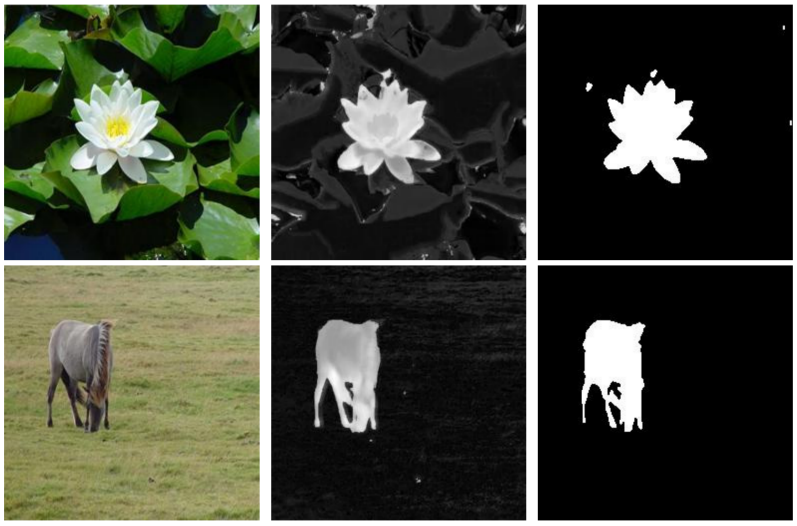
3.4.1 Automatic Graphcut Segmentation

Segmenting foreground objects out of the image is an important task in computer vision research. The graphcut based techniques which are adopted by popular commercial softwares outperformed than many other methods. By scattering the seeds labeled as foreground and background, the graphic model eventually determines the attributes of the pixels and produces the binary masks. The significant drawback of this segmentation method is that it should manually choose the initial seeds to implement the cutting process. Introducing the saliency map we got, the seeds selection procedure is transferable to saliency guided strategy. The most saliency pixels are automatically chosen as the foreground, meanwhile, the corresponding pixels with low saliency values are tagged as background. We then experimented the growcut algorithm following Vezhnevets and Konouchine (2005) to generate the saliency based segmentation on various images. Our saliency maps help the



(a) (b) (c) (d) (e) (f) (g)

Figure 3.14: Saliency maps comparison. From the top row: original input images, saliency maps generated by IT, SR, MZ, GB, CA, AC, LC, FT, HC, RC, LR and our proposed method SMAP. Columns (a)(b) demonstrate the color independent attribute which means the saliency map does not rely on color. Column(c) demonstrates the color uniqueness in multi-color environment. (d)(e) illustrate the scenario that the background contains texture information. Notice the red flower held by the toy bear in column(f), the proposed SMAP method is the only one detected efficiently the red part as the saliency part. In column (g), even in this extreme case, the gull is still detectable using the proposed algorithm.

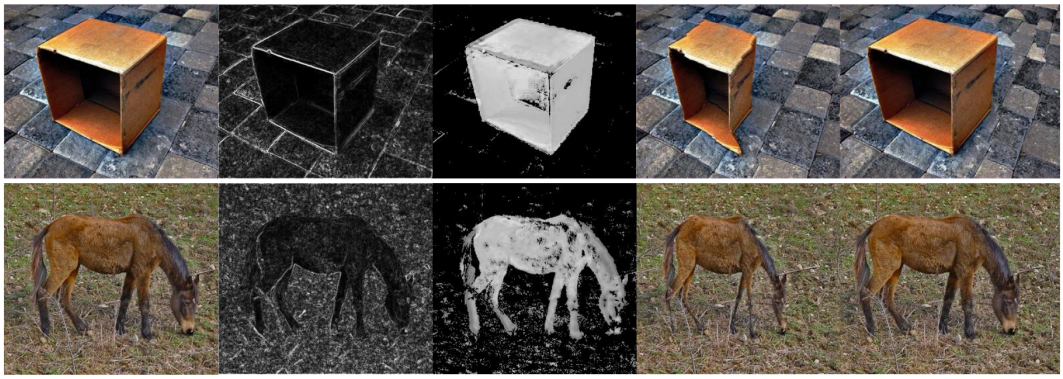


growcut method with a high confidence to segment single/multiple object(s) in clutter environment.

3.4.2 Image Retargeting

Image retargeting technique is developed for resizing images that is adaptive to the image content. It functions by establishing a number of non-informative seams and remove them to shrink the size of image. Therefore, image retargeting performance greatly relies on accurate saliency map generation algorithms to locate 'important' and 'boring' bits on image. The reducible seams on dominant objects can be effectively avoided by adopting our saliency detection algorithm. We performed our saliency maps in image carving method proposed by Avidan and Shamir (2007). In the original algorithm, the seams generated by calculating the distortion energy on relative non-saliency regions other than the featured pixels. However, the energy map used cannot be guaranteed to be uniformly distributed on the whole saliency objects. The reducible seams generate not only in the background but also across the targets and resulted in a distortion in the interesting objects. Comparative analysis on the proposed saliency map preserves both the clear edges and highlights the content of the completed saliency objects. Thus the outcomes of image resizing are smooth and

uniform. In Fig.3.15, it is clearly to see the difference between the proposed saliency map and normal resizing scheme.



(a) (b) (c) (d) (e)

Figure 3.15: Saliency map assisted image retargeting. (a) original images; (b) default energy map by algorithm Avidan and Shamir (2007); (c) saliency map generated by the SMAP; (d) retargeting results by Avidan and Shamir (2007); (e) retargeting results by the SMAP algorithm.

3.4.3 Scene Depth Effect on Commercial DC

Instamatic camera is a popular digital device in modern household and widely used in smartphones. For its convenient property, more than 30% digital photos are produced by instamatic cameras. However, limited by its small CCD and aperture range, the scence depth effect cannot be rendered directly from the pictures. Inspired from the focusing and aperture tuning functions of the digital single lens reflex camera (DSLR), the scene depth effect can be emulated by keeping the focus on saliency objects and blurring the unimportant background from normal pictures. We designed the algorithm which applied the Gaussian blur function based on saliency map. The higher saliency value of a pixel, the lower strength of blurring effect it acquired. The simple idea generates the photo with synthesized scene depth effect comparable to the realistic photos produced by DSLR. (see Fig.3.16)

From the demonstrated images, the focused figures kept original resolutions while the background patches were blurred at a certain degree to emulate the depth effect.

Figure 3.16: Saliency map assisted scene depth effect rendering.

3.5 Conclusion

In this chapter, we performed a three-level saliency detection strategy, SMAP, to analyze the saliency attribute of the images. It was implemented in a simple structure which combines the local contrast technique based on bio-inspired attention feature and the global color distribution constraint. From the experimental results, we observed that the proposed approach fully satisfies the criteria of biological observation on human vision and related application requirements. The proposed technique emphasizes on segmentation enhancement application in the clutter environment with full resolution requirement. The quantitative precision-and-recall curves illustrated that our approach outperforms the state-of-the-art works.

Chapter 4

Object Recognition via $L_{1/2}$ Norm Regularized ADN

4.1 Introduction

Object recognition is a process for identifying a specific object in a digital image or video. Every day we discover and recognize a multitude of familiar and novel objects. As the most fundamental capability, human can do this with little effort, despite the fact that these objects may vary somewhat in form, color, texture, etc. Objects are recognized from many different vantage points (from the front, side, or back), in many different places, and in different sizes. Objects can even be recognized when they are partially obstructed from view. To be able to accurately recognize the subtle discriminants in the image would benefit a wide range of applications, for instance, video stabilization, automated vehicle parking systems, and cell counting in bio-imaging. Object recognition algorithms rely on matching, learning, or pattern recognition algorithms using appearance-based or feature-based techniques. Common techniques include edges, gradients, Histogram of Oriented Gradients (HOG) Dalal and Triggs (2005), Haar wavelets Chen and Hsiao (1997), and local binary patterns Ahonen et al. (2006); Ojala et al. (2002). Understanding

its complex origin and processing mechanism, automatic object recognition remains a great challenge in computer vision research.

To capture the subtle feature for automatic object recognition, we seek to extract efficient representation from the raw images. To date, there have been two branches of works in this area, roughly divided by the type of features extracted: feature-based scheme and appearance-based scheme. The feature-based scheme aims to propose a search that is used to find feasible matches between object features and image features. The recognition relies on accurate and reliable feature detection, tracking and geometric constraint. Comparably, in the appearance feature based approach, the local features are often applied to model the appearance of the recognizing objects in terms of feature descriptors. Usually, the edge, corner, texture and color histogram features play the dominant role for object recognition. However, without the proper post-process like feature selection, the generated feature contains too much redundancy and thus degrades the performance in recognition.

To address the problems in object recognition, we investigate into the hierarchical structure of facial expression through the adaptive deconvolutional network (ADN) Zeiler et al. (2011). The deep structured ADN is firstly proposed for objects categorization problem. The most important advantage of ADN is the unsupervised feature learning capability. Since the reconstruction constraint is applied layer-wise on the stacked network, the extracted feature set is complete and hierarchical. The de-convolution process in the intra-layer and the max pooling in the inter-layer enable the network to exploit the most representative features. In this paper, we reformulate the original deep ADN model by replacing the L_1 norm with the proposed $L_{1/2}$ norm. In addition, the ADN recognition architecture is redesigned to be adaptive for FER. According to the existing literatures, the state-of-the-art results reported in object recognition are mostly involved with a handcrafted feature selection or supervised labeling. Without such a human guided pre-process, the proposed $L_{1/2}$ norm regularized ADN is advantageous in terms of spontaneously constructing the hierarchical features of the object. Our work demonstrates that, the proposed

ADN can derive more compact features leading to more robust and reliable object recognition performance, as validated by the comprehensive experiments in the large dataset.

4.2 Feature Learning Approach: Adaptive Deconvolutional Network

In this section, we first introduce the hierarchical feature learning framework based on the Adaptive Deconvolutional Network (ADN) that decomposes the input image into deep structured feature sets. We then describe the proposed $L_{1/2}$ norm regularization as the deconvolutional sparsity constraint that exploits the subtle feature in object description to facilitate more accurate recognition. Based on the dense feature vector produced from the previous learning and inference procedures, we apply the SVM classifier for object recognition.

4.2.1 Feature Learning through Adaptive Deconvolutional Network

The ADN is a multi-layer trainable architecture that can learn hierarchical set of feature maps from input images Zeiler and Fergus (2014); Zeiler et al. (2011); Jamieson et al. (2012). The whole structure consists of the sparse constrained convolutional layers (named 'deconvolutional') and max-pooling. In each of the deconvolutional layers, l, the input image y is decomposed into K_l feature maps $z_l = \{z_{k,l} | k = 1, \cdots, K_l\}$ convolved by the learnt K_l filters $f_l = \{f_{k,l} | k = 1, \cdots, K_l\}$,

$$\hat{y}_l = \sum_{k=1}^{K_l} z_{k,l} \otimes f_{k,l} \tag{4.1}$$

For layer l, the training procedure minimizes the training cost function,

$$C_l(y) = \frac{\lambda_1}{2}\|\hat{y}_l - y\|_2^2 + \sum_{k=1}^{K_l} |z_{k,l}|_1 \tag{4.2}$$

where the first term is the image reconstruction cost, and the second term is the L_1 norm sparse penalty applied on the learnt feature maps $z_{k,l}$, aiming to discourage changes in the features associated with small changes in the input images.

At the top of each deconvolutional layer, the 3D max-pooling is performed to shrink the feature maps by pooling the local maximal pixel values within the 2D image field as well as the neighboring channels. During the max-pooling, the characteristic process 'switching' is also performed. The locations of the pooled maxima are stored by the switches. Taking the switch s as an output augment, the pooling operation is treated as a linear process: $[p, s] = P(z)$, where the specified elements in z are copied to p and their locations are recorded in s. With such setting, the unpooling operation U_s is also a linear process, where elements in p are copied to the reconstructed feature map $\hat{z}$. The remain elements in $\hat{z}$ are set to zero: $\hat{z} = U_s p$. $U_s = P^T$ is captured.

The *inference* from features on l layer is defined with *reconstruction* process R_l,

$$\hat{y}_l = F_1 U_{s1} F_2 U_{s2} ... F_l z_l = R_l z_l \tag{4.3}$$

F_l contains the convolutional results between filters f_l^k and feature maps z_l^k. In learning on l^{th} layer, we firstly infer the reconstructed $\hat{y}_l$ according to Eq. (4.3), and then compute filters f_l by optimizing the objective function in Eq. (4.2).

The multi-layer learning strategy remains the same as the aforementioned method , except that the number of feature maps increases. By max-pooling and increasing the number of feature maps, the sizes of receptive fields in the feature maps have been changed, resulting in the formulation of hierarchical features. Details on the filter learning and feature map inference procedures are illustrated in Fig. 4.1, and the mathematical details can be found in Zeiler et al. (2011).

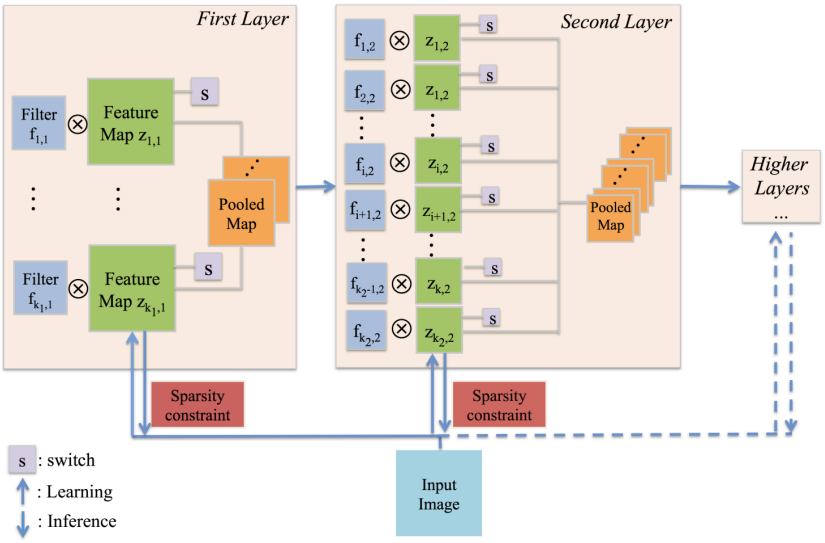


Figure 4.1: Illustration of the Adaptive Deconvolutional Network (first two layers).

4.2.2 $L_{1/2}$ Norm Regularization on Feature Learning

Recently, sparsity has become a necessary requirement in both statistics and optimization tasks in order to control the dimension complexity for many applications Guo and Qi (2013). It has been shown, through biological experiments that sparsity is natural process in the entire hierarchical processing of visual information Bengio (2012). For example, the L_0 norm regularization on representation learning encourages to learn sparse and discriminant features. In the work of Donoho (2006), the L_1 norm regularization has been proved to have the equivalent ability as the L_0 constraint on sparse signal reconstruction. However, when we consider the unsupervised feature learning task in FER, the capability to decouple tightly mixed factors of variation underlying expression, facial morphology and nuisances is highly desired. L_1 norm penalty, as a convex optimization, is suffering in enforcing further sparsity and often leading to an over-penalized regularization Xu et al. (2012).

To exploit the more discriminant features, we adopt $L_{1/2}$ norm regularization as the sparsity penalty on the layer-wise learning during ADN model construction. Thus, the per-layer cost function is redefined as,

$$C_l(y) = \frac{\lambda_1}{2}\|\hat{y}_l - y\|_2^2 + \lambda \cdot \sum_{k=1}^{K_l} \|z_{k,l}\|_{1/2} \tag{4.4}$$

where, $\|z_{k,l}\|_{1/2}$ represents the $L_{1/2}$ quasi-norm, and λ is the regularization parameter to coordinate the strength of model accuracy and sparse penalty. The $L_{1/2}$ norm regularization is a natural improvement to convert FER feature extraction into a non-convex and non-smooth optimization, which is more realistic in real-world scenario.

The sparsity of feature maps z_l is learned by applying hard thresholding in ISTA Beck and Teboulle (2009) iterations at the inference phase. The computation scheme contains iterative gradient and shrinkage steps. In gradient step, the reconstruction error is calculated with respect to z_l, and then the gradient g_l is defined as $g_l = R_l^T(R_l z_l - y)$. Once g_l is computed, z_l is updated by $z_l = z_l - \lambda_l \beta_l g_l$, where β_l parameterizes the gradient step size. In shrinkage step, the per-element shrinkage operation is added to enforce the sparsity by

$$z_l = \begin{cases} z_l - sgn(z_l)\beta_l & |z_l| > \beta_l \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

Inspired by the ISTA, we adopt the half thresholding Xu et al. (2012) to solve the $L_{1/2}$ norm regularization by shrinking z_l in terms of

$$z_l = \begin{cases} \frac{2}{3} z_l (1 + \cos(\frac{2}{3}\pi - \frac{4}{3}\varphi(z_l))) & |z_l| > \frac{\sqrt[3]{54}}{4}(2\beta_l)^{\frac{2}{3}} \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

where $\varphi(z_l) = \arccos(\frac{\beta_l}{4}(\frac{|z_l|}{3})^{-\frac{3}{2}})$. The convergence of the solver is proved in Xu et al. (2012).

The chosen $L_{1/2}$ norm penalty is not an ad-hoc constraint for ADN framework. By introducing the sparse constraint, we are expecting to decompose images into patches with oriented gradient feature in different scales. These features are called Gabor-like features. The final recognition can be explained as the descriptor matching with a

designed distance metric. Recently, the related research reveals that the statistics of such gradient based feature matching derives a heavy-tailed distribution Jia and Darrell (2011). Considering such a prior, $L_{1/2}$ norm regularization performs better in handling signal reconstruction with a heavy-tailed distribution which has been proved in theory Xu et al. (2012).

4.2.3 Feature Vector Formulation and Classification

The ADN model learns the hierarchical image features in an unsupervised way. In other words, given the input images, the facial expression is automatically decomposed into multi-layer feature sets on the layers with the fixed filters. However, the switches between different images are not with the same configurations. That means, within the same expression class, the image decompositions share similar features, but their reconstructions are quite different from each other. Thus, for expression recognition purpose, it is problematic to use the learned feature maps directly. An alternative method is to construct the feature vector by selecting the largest M activations from the top layer and re-project them to the first layer and use the reconstructed images to represent the input. The reconstructed images on the first layer are subdivided into equal-size small patches with spaced pixel shifts ('stride's). Within each patch, the max-pooling is processed to generate a single value to represent the whole patch. The down sampled images from different channels are reshaped, stitched and concatenated into a vector in our work, which is the final representation for an input image used for recognition. The entire process to construct the feature vector is illustrated in Fig. 4.2.

We add a supervised classifier upon the learned features to implement the classification. After the feature vectors are generated, we apply the SVM classifier with the RBF kernel to expression recognition. All the classifications are based on 4-fold cross validation. Due to the dense sampling in the aforementioned feature vector

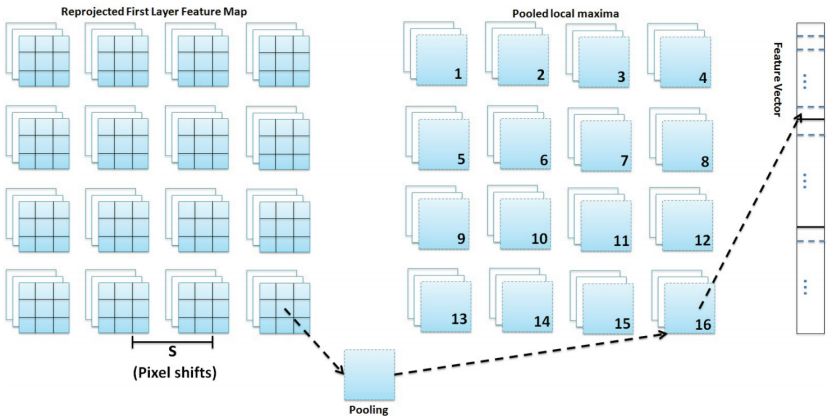


Figure 4.2: Feature vector formulation. The input is the projected first layer feature maps.

formulation, the resulting feature vectors are of high dimension. Before we feed the feature vectors into the classifier, we also apply PCA to reduce the dimensionality.

4.3 Object Recognition via $L_{1/2}$ Norm Regularized ADN: Evaluation

The object recognition experiment is carried on with the same MSRA saliency dataset. There are totally 1000 images. The input image is not the raw image but the image patch with detected salient objects. Since the dataset is not naturally designed for object recognition, the image annotation is manually made. The whole dataset is labelled as Animal, Bird, Building, Car, Plant, Human, Traffic Sign and Miscellaneous class. To avoid the trivial affection from miscellaneous class, in the recognition experiment, we use the images of first 7 classes but remove the miscellaneous class. For testing, the MSRA saliency dataset B is used. There are totally 5000 images (1000 duplicates to the original MSRA saliency dataset). Following the same criteria, totally 585 testing images in aforementioned 7 classes are selected to formulate the testing set. The statistics of the testing dataset is list in Table. 4.1. The dataset is demonstrated in Fig. 4.3.

Table 4.1: Statistics of the testing dataset from MSRA dataset B

Total	Animal	Bird	Building	Car	Plant	Human	Traffic Sign
585	**87**	**74**	**68**	**72**	**118**	**85**	**81**

Figure 4.3: Demonstrations of MSRA saliency dataset for object recognition. From top row to the bottom: Animal, Bird, Building, Car, Plant, Human and Traffic Sign. The demonstrated images are saliency detection results. All the images are normalized into the size of 256×256.

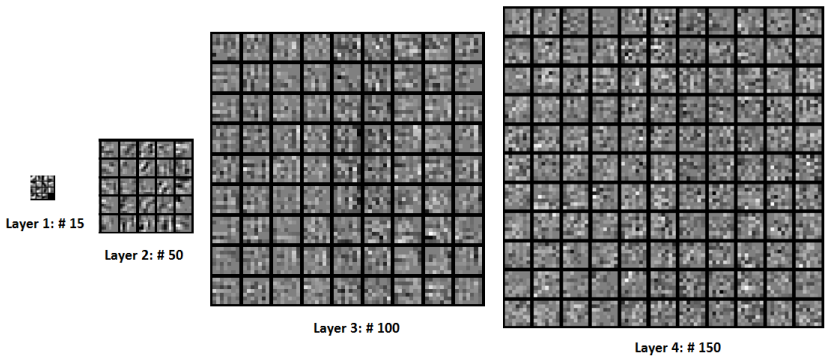


Figure 4.4: Learned filter kernels after training using 1000 saliency patches from MSRA dataset. The numbers of kernels in each layer are 15, 50, 100 and 150 respectively. Each of them is of size 7×7.

Pre-processing: Each image is converted to gray-scale and resized to 256×256 (Bicubic interpolation). Local subtractive and divisive normalization (i.e. the patch around each pixel should have zero mean and unit norm) is applied using a 1313 Gaussian filter with $= 5$. Since the images are various in illumination condition, the pre-processing helps to reduce the negative effect from the lighting.

Model architecture: We use a 4 layer model, with 7×7 filters, and E = 10 epochs of training. From layer 1 to layer 4, the numbers of filter kernel are 15, 50, 100 and 150. Benefiting from the efficient inference scheme, the proposed ADN is able to handle with many more feature maps and more data than conventional approaches. By the 4th layer, the receptive field of each feature map element covers the majority part of the image (189×189), making it suitable for the novel feature extraction considering the representative and discriminant attributes.

Timings: With 585 training images and E = 10 epochs, it takes around 26 hours to train the entire 4 layer model using a MATLAB implementation on intel i-5 CPU (laptop environment without GPU acceleration). For inference, a single epoch suffices with 10 ISTA iterations at each layer. The total inference time per image is 0.79 sec. The learned filter kernels are demonstrated in Fig. 4.4. The learned feature maps are also visualized in Fig. 4.5.

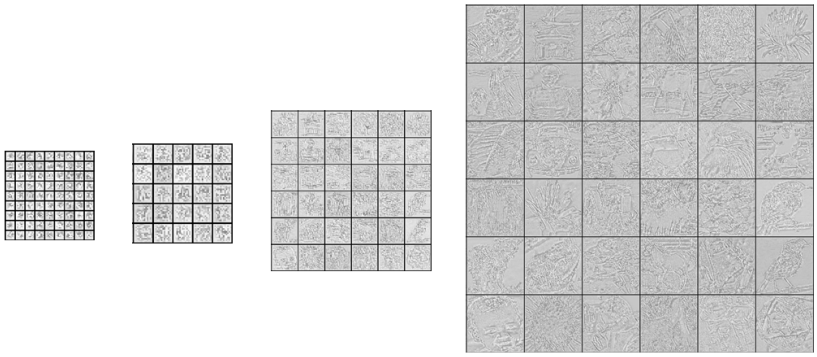


Figure 4.5: Learned feature maps for each layer. The leftmost are feature maps from layer 1 and the rightmost are feature maps from layer 4. Clearly, the fourth layer feature maps have already depicted object contours and detailed structures.

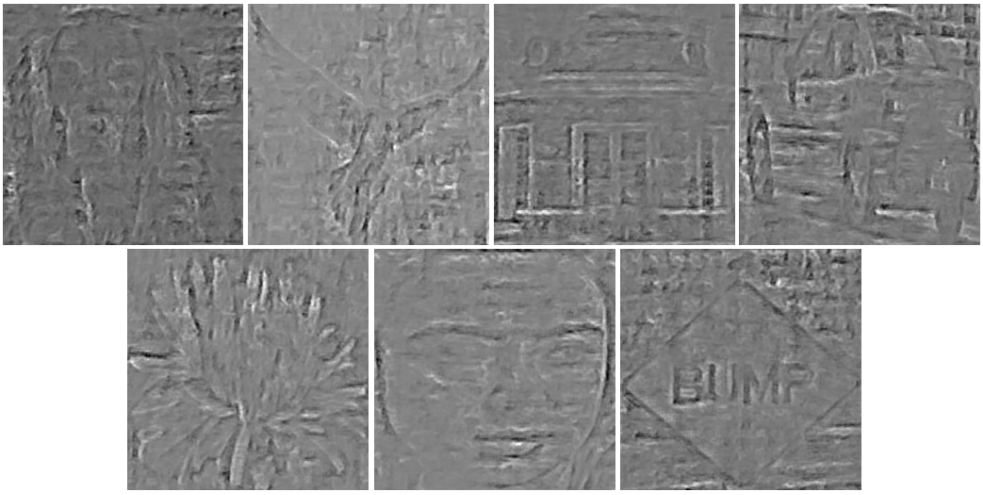


Figure 4.6: Reconstructed images on image layer with $M = 100$ activations in the fourth layer.

By utilizing the training and feature vector formulation strategy described in previous section, we choose $M = 100$ at the fourth layer and re-project them back to the pixel layer. Several reconstructed images in testing dataset are demonstrated in Fig. 4.6. We trained the SVM classifier with RBF kernel on the generated feature vector and calculated the recognition accuracy. We use 2-fold cross-validation in the testing stage. Recognition accuracies are reported in Table. 4.2.

Appplying the SVM classifier to layer 4 features from the proposed ADN structure produces the best results. Although the accuracy is lower than the convolutional

Table 4.2: Recognition performance on MSRA saliency dataset. The comparison approaches include PCA, SIFT and 5 layer CNN structure.

Proposed ADN + layer 2	64.6 ± 2.7%
Proposed ADN + layer 3	70.6 ± 3.2%
Proposed ADN + layer 4	**81.3 ± 2.4%**
Original ADN + L1 norm	72.3%
PCA + SVM	63.5%
SIFT + SVM	69.8%
CNN + 5 layers	**88.4%**

Table 4.3: Performance comparison between saliency detection results, ground truth object patch and entire image, as the training and testing inputs.

Proposed ADN + saliency detection	**81.3%**
Proposed ADN + ground truth patch	**84.0%**
Proposed ADN + entire image	68.2%

neural network with the same task, it is acceptable that CNN is supervised training. It asks for extra labeling information to fine tune the network parameters so as the classifier. The proposed model received a gain of 11% over SIFT feature based approach indeed encourages the research.

Noticed that, the input training and testing images are from the saliency detection results which contains imperfect detection. Another experiment is conducted to evaluate the performance of the proposed model on ground truth object patches and the raw images without detection. It reflects the efficiency of the proposed ADN so as to the complexity of the dataset. The final recognition results are illustrated in Table. 4.3.

More detailed analysis of the $L_{1/2}$ norm regularized ADN can be found in the following section with a case study on facial expression recognition.

4.4 Case Study: Facial Expression Recognition via $L_{1/2}$ Norm Regularized ADN

To evaluate the strength of the proposed ADN, we also conduct a case study by applying the ADN structure on facial image data and implementing the facial expression recognition.

From the psychology society, the recently FACES database Ebner et al. (2010) provides a relatively large expression library. The FACES database contains two sets with the same 171 individuals. Each person performs six expressions with the frontal view position. Each expression is performed twice and separated into two sets. So we have 2052 images in total.

Another labeled expression database is the Lifespan database Minear and Park (2004). It is more challenging considering its diversity in races of performers, age span, slight pose variations and uneven numbers of images for each expression. We take only two types of expressions 'Happy' and 'Neutral' from this database for evaluation usage (comparable to the existing work). There are 590 neutral face images and 254 happy face images in the database.

FER-2013 is one of the largest publicly accessible facial expression datasets Goodfellow et al. (2013). The entire data consist of 28709 training images with unified 48×48 resolution under 7 different types of expression. The testing set is of 3589 images. We conduct the classification task on the dataset to validate the robustness of the extracted features in the proposed unsupervised manner. The examples of images from the dataset are shown in Fig. 6.6.

To perform a comprehensive investigation on the performance of the $L_{1/2}$ norm regularized ADN, we design four experiments using the FACES and Lifespan facial expression databases. In the first two experiments, we retrieve the layer-wise hierarchical features learnt by the proposed deep network and inspect the effect of $L_{1/2}$ norm regularization by comparing the expression recognition accuracy with that using the L_1 norm regularized network . The third experiment evaluates the adaptation

Figure 4.7: Expression databases illustration. The first row contains the six expressions from FACES (Happy, Angry, Fearful, Sad, Disgusted and Neutral). The second row is the Lifespan database with two expressions. All the images are cropped based on region of interest for further usage. The last two rows contain the seven expressions from FER-2013 (Angry, Disgust, Fear, Happiness, Sadness, Surprise and Neutral).

of the ADN with respect to different image resolutions. In the fourth experiment, we apply the model learnt from the FACES database on the Lifespan database to validate the transfer learning capability of the proposed framework.

4.4.1 Visualization of the Learnt ADN and Layer-wise Comparison for Expression Recognition

The original FACES images are of the size 2500×2500. We crop out the region of interest (ROI) containing the majority of face (from forehead to chin, ear to ear) and downsample them to 128×128 to save the computation time. We randomly select 504 images from the whole database and use them as the training set. The remaining 1548 images are formulated as the testing set. There is no identity overlapping between the training and testing sets, which means the images from the same person can be either in the training group or the testing group, but not both.

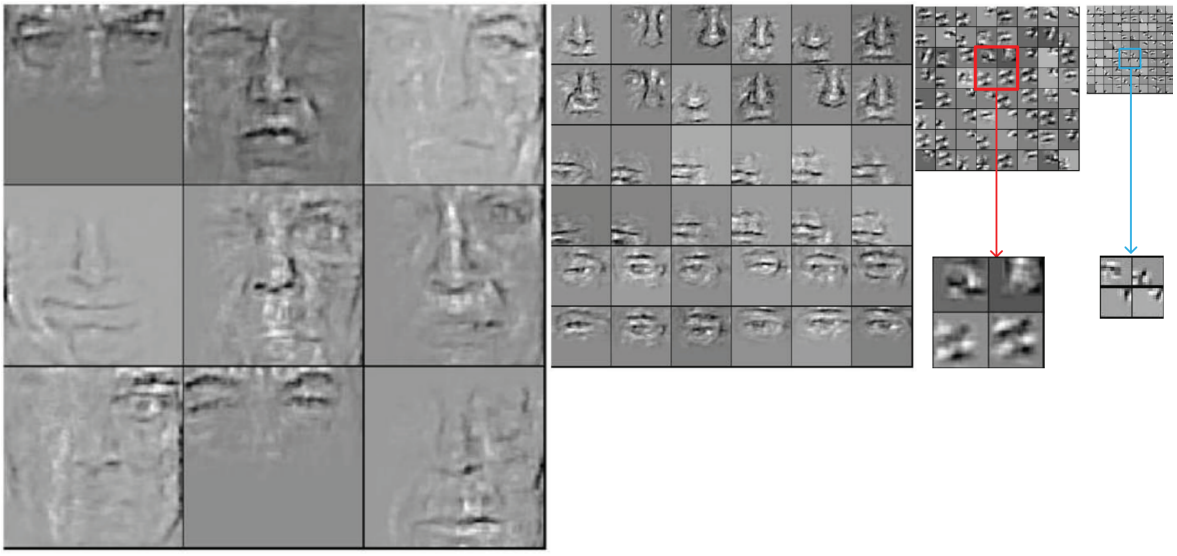


Figure 4.8: Hierarchical features learnt by the proposed ADN architecture. Feature generated by projecting the largest one activation from each layer back to the pixel space. From left, features learnt by layer 4 to layer 1. The activation from layer 4 has the receptive field covered the entire face. In the 3^{rd} layer, features are acquired at the facial parts level (nose, eyes, mouse, etc.). Features in 2^{nd} layer are mostly basic junction parts. In the 1^{st} layer, the primitive level Gabor-like features are learnt. The four-layer feature sets form the feature hierarchy. Noted that features are not in the original scale.

A four-layer ADN is trained in our experiments with the 7×7 filters in each layer. The general parameter configurations and statistics are shown in Table 4.4. Other latent parameters have the same setting as in Zeiler et al. (2011). These settings remain the same for the following experiments until changes are mentioned.

By feeding the training set into the ADN network, the top-down hierarchical representation is learned. As the pooling process being applied, the feature maps z_l shrinks in size but increases in numbers and size of receptive fields. More and more distinct features are learnt gradually. Different from other deep feature hierarchies, in the proposed ADN model, the features at all layers can be visualized with a clear semantic meaning. From the first layer, the face image is decomposed into facial parts (facelets), followed by layers subdividing the features in smaller scales that form the junctions, curves, edges till the oriented Gabor-like features in the fourth

layer. The learnt hierarchical features are illustrated in Fig. 4.8. At the 4^{th} layer, the receptive field of each feature map element covers the whole image range. The largest activations picked from this layer reflect the strongest response towards the input stimuli, and the re-projection back to the first layer feature maps with these activations are deemed as the appropriate features for expression recognition.

Table 4.4: General L_1 and $L_{1/2}$ norm regularized ADN parameter setting and layer-wised recognition performance. The last two rows contain the recognition accuracies for L_1-ADN and $L_{1/2}$-ADN respectively.

Property	**Layer 1**	**Layer 2**	**Layer 3**	**Layer 4**
# of F-Maps	15	50	100	150
Pooling Size	$3 \times 3 \times 3$	$3 \times 3 \times 2$	$3 \times 3 \times 2$	$3 \times 3 \times 2$
Recep Field	7×7	21×21	63×63	189×189
Feat Dims	134×134	51×51	23×23	14×14
L_1**-ADN**	58.70%	61.06%	68.31%	70.59%
$L_{1/2}$**-ADN**	74.13%	69.22%	64.89%	**81.70%**

Considering the expression recognition task, the straightforward question is that at which level the expression manifold is best represented. To inspect this, we test the trained ADN with the testing dataset, and construct all-layer features. Then the feature vectors from each layer are extracted and used for the SVM classifiers following the method* discussed in Section 4.2.3. The classification results and comparison are shown in Table 4.4. The feature using top layer activations produces the highest recognition accuracy. As a comparison, in Guo et al. (2013) which uses the handcrafted Gabor features on the same database, it only gives 69.32% in accuracy, as shown in Table 4.5. At this point, we have successfully constructed the ADN based deep network and applied it to the expression recognition task. We also conducted the experiments on the state-of-the-art methods with the same dataset. The final results are reported in Table 4.5. As an unsupervised deep hierarchy, the proposed architecture performs only two major computations: intra-layer convolution and intra-layer max-pooling, but gains the capability to extract the expression related

*To generate the feature vector, pixel shift is set to 4, patch size is equal to 4×4. There is no overlap between patches. PCA is set to cover 95% of the information. The number of top largest activations selected is set to $M = 100$

Figure 4.9: Demonstrations of the pooling locations on the images. The red blocks represent the pooling position at one channel. Notice that, most of the pooling position are coincident to the local landmarks on the face.

components from the complex facial image space and receives competitive results in FER tasks.

ADN belongs to the family of bio-inspired models which also include the famous works CNN LeCun et al. (2010) and HMAX Serre et al. (2005). In the HMAX model, a hierarchy of increasingly complex features are generated by alternating template matching and max-pooling. In particular, at its $S1$ layer, the input image is firstly convolved with the Gabor filter banks to closely mimic the visual cortical processing. We receive the same feature maps in the higher layer and this kind of Gabor-like features is coincidentally used in the state-of-the-art expression recognition works and received the highest classification accuracy in the reports. The max-pooling is also shared in HMAX and ADN, leading to increased invariance to distortions. Facilitated with the 'switch' setting, when we re-project the features from top layer back to the image space, we are able to locate the pooling positions precisely. Visualizing these pooling locations, we find that most of these points align with the key feature points (FPs) representing the landmarks on the face Bettadapura (2012). The ADN captures both the local appearance and globally geometric shape in its simple operations while at the same time to strengthen itself in expression recognition capability.

Table 4.5: FER accuracy comparison. For LDA Yu and Yang (2001) (Linear Discriminant Analysis), 504 images are used for training, the rest are used as testing samples; for RI-LBP Shan et al. (2009) (Rotation-Invariant LBP), we use one-vs-all classification scheme and SVM as the classifier; for CNN Phung and Bouzerdoum (2009), we use one-vs-all classification scheme and percepton as the classifier.

Approaches	Recognition Accuracy
LDA	29.54%
Gabor ?	69.32%
RI-LBP	79.20%
CNN	80.95%
$L_{1/2}$**-ADN (Layer 4)**	**81.70%**

Table 4.6: Recognition accuracies comparison based on FER-2013.

Methods	**Recognition Accuracy**
Human Accuracy	$65 \pm 5\%$
DLSVM L2 Tang (2013)	71.16%
L1/2-ADN (Layer 3)	**61.48%**

4.4.2 The Robustness of The Unsupervised Feature

The robustness of the extracted feature against pose, illumination and occlusion variations is an essential factor in the real world FER task. We conduct the experiment to validate such property over the FER-2013 dataset. Since the image resolution is 48×48, we use a different configuration of the network by reducing the number of layers to 3, meanwhile, fixing the kernel size of all the filters and pooling operation. Consequently, the feature map dimensions are changed into 54×54, 24×24 and 14×14 respectively in each layer. We keep the same processes for the feature vector generation and recognition using SVM. The learnt filter kernels and features are shown in Fig. 4.10. The final FER accuracy is reported in Table. 4.6.

According to Goodfellow et al. (2013), the human recognition accuracy on FER-2013 is $65 \pm 5\%$ due to the pose, illumination and occlusion variations included in the dataset. The winning solutions in FER-2013 challenge are all supervised schemes or using handcrafted features Goodfellow et al. (2013). We received encouraging results in terms of the unsupervised learning approach embedded in the proposed $L_{1/2}$-ADN

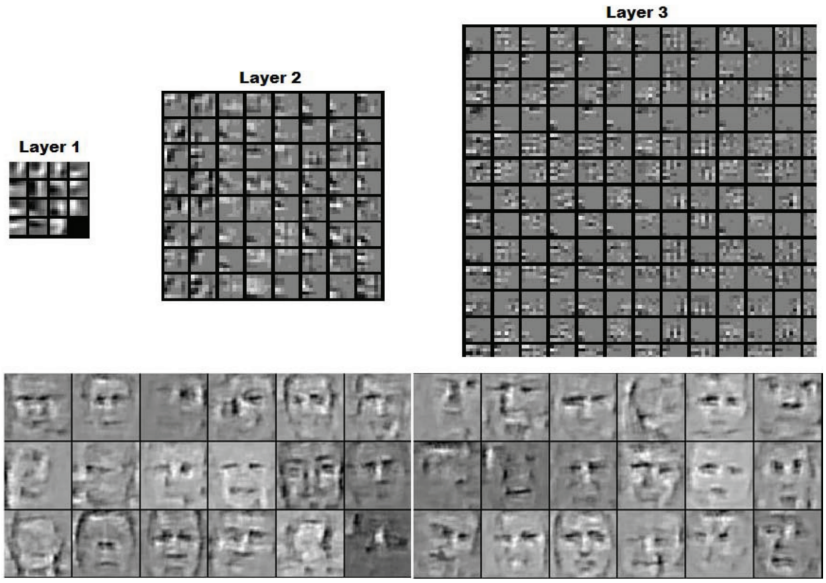


Figure 4.10: Demonstrations of the learnt filter kernels and projected features from 3^{rd} layer activations to the image space on FER-2013 dataset. We have 15, 50 and 100 filters on each layer.

for FER task. The layerwise convolution and the pooling indeed assist in capturing subtle semantic features from the facial images.

4.4.3 The Role of $L_{1/2}$ Norm Regularization

Sparsity constraint is deployed to encourage learning of distinctive features in the representation learning. Mostly, L_1 norm regularization is adopted as a proxy for optimizing L_0 sparsity. However, L_1 regularization cannot give out the sparsest solution for certain data distributions, i.e. uniform distribution, heavy-tailed distribution Xu et al. (2012). We introduce the $L_{1/2}$ norm regularization with iterative solver to the ADN framework aiming to exploit the optimized features in the expression manifold. The quantitatively analysis of the $L_{1/2}$ norm regularization is conducted from two aspects. For comparison purpose, we train two ADN models with the L_1 regularization prior and the proposed $L_{1/2}$ regularized prior respectively with the same network settings. The layer-wise recognition results are shown in Table 4.4. We evaluate the facial image reconstruction capability and expression recognition

Figure 4.11: Demonstrations of the image reconstruction using 4^{th} layer feature activations. From left: original input image (gray value), reconstruction with $L_{1/2}$ norm regularized ADN and the reconstruction with L_1 norm regularized ADN. From the figure, the left side nasolabial fold cannot be well reconstructed in the L_1 norm regularized ADN. The MSE is reported in Table 4.7.

accuracy with the learnt feature maps at the 4^{th} layer of the two regularizations. The reconstruction error is measured as MSE.

Table 4.7: $L_{1/2}$ norm and L_1 norm regularization comparison in image reconstruction

Experiments	**Recon Error (MSE)**
L_1-ADN (Layer 4)	**1.8311e-04**
$L_{1/2}$-ADN (Layer 4)	**6.1035e-05**

The reconstruction results in Table 4.7 suggest that the ADN regularized by the $L_{1/2}$ norm constraint is more efficient in distinctive feature learning to decouple the expression manifold factors from the image space. Both the L_1 norm penalty and $L_{1/2}$ norm constraint force the representation learner to generate the Gabor-like features at the higher layers. The layer-wise recognition results are shown in Table 4.4. However, from our experimental results, $L_{1/2}$ norm regularization indeed produces more discriminant feature for expression representation. As reported in the original ADN paper Zeiler et al. (2011), one of the reasons to project the top layer activations back to the 1^{st} layer is that, the feature maps perform similarly as the dense SIFT feature descriptors. The visualization of 4^{th} layer features in Fig 4.8 clearly demonstrates the property of the generated features is oriented gradient based.

We conduct the statistics on the feature maps by calculating the histogram of derivatives $\nabla_x F$ and $\nabla_y F$, where F represents the 4^{th} layer features. The generated curves are consistent with the assumption that the features are highly non-Gaussion,

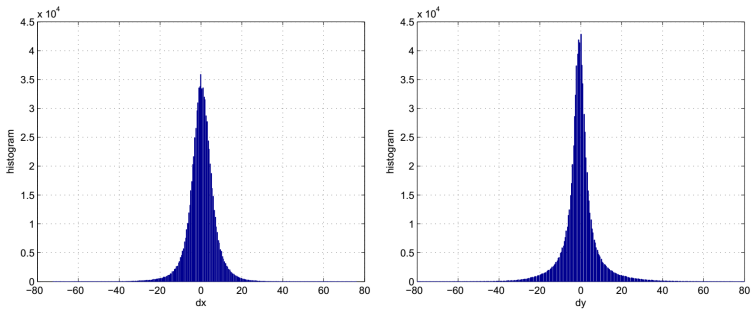


Figure 4.12: Histogram of $\nabla_x F$ (left) and $\nabla_y F$ (right) by accumulated 50 facial image feature maps on 4^{th} layer.

but kurtotic with a heavy tail, as shown in Fig. 4.12. Thus, the $L_{1/2}$ norm constraint is naturally validated to be more efficient in FER tasks.

4.4.4 Effect of Multi-resolution

Although the ADN enables the efficient learning of multi-scale features to represent the input images, as a hyper parameter, the filter size is determined empirically in all related literatures Zeiler and Fergus (2014); Zeiler et al. (2011). The relationship between the input image resolution and the deep structured filter size is rarely studied.

In this experiment, we investigate into the effect of input image resolution on the expression feature extraction with fixed filter size. The input images are re-scaled to 64×64 and 256×256 separately. The stacked filters in the ADN keep their original size. Then the dimension of the learnt feature maps and their receptive fields at each layer are changed. For the 64×64 dimensional images, the newly required receptive field at the 3^{rd} layer has already covered the whole image range. Comparatively, for the 256×256 sized image, the receptive field in the top layer shrinks to the limited scope of the image. The top layer features degenerate to local features. Other network configurations keep the same as the previous experiment, and we show the expression recognition results in Table 4.8.

When we look close to the recognition results, although the 256×256 resolution images contain more facial details, instead, the 128×128 images receive the highest

Table 4.8: Comparison between multiple input resolutions with $L_{1/2}$-ADN and 4^{th} layer features

Property	**Feature Map Dims (4th Layer)**	**Recog Acc**
Res. 64×64	12×12	69.32%
Res. 128×128	14×14	**81.70%**
Res. 256×256	19×19	57.76%

recognition accuracy. In the learnt feature hierarchy, the first layer feature maps are obtained by convolving the image patch-wise with the fixed sized filters. The size of filters determines the strength scale. Combined with pooling size, it also determines the receptive field of one element in each layer, and so, define the type of feature is a global one or not. The compatible correlation between the filter size and image resolution results in both the best local representation and global abstraction for ADN in face image feature learning. Such a corporative setting is extremely desirable in the challenging tasks, such as FER. Based on our observation, the filters sized to 7×7 are suitable to measure the scale of the major landmarks (eyes: 20 to 25 pixels in length; nose: 50 pixels in length; mouse: 60 pixels in length) and beneficial to obtain the global geometry of the face with 128×128 resolution inputs.

4.4.5 Transfer Learning: Feature Adaptation

A good representation learning algorithm is expected to exploit the commonalities between different learning domains in order to share and transfer learnt knowledge across databases Bengio (2012). We hypothesize the $L_{1/2}$ norm regularized ADN has such advantage in feature adaptation because of its unique learning manner that designed to learn the distinctive features from facial images. By using the filters trained on FACES and applying them to classify the Lifespan database, we validate the generalization attribute of the proposed ADN model. Particularly, in this experiment, we preprocess the Lifespan images the same way as we did on the FACES images (crop ROI and rescale into 128×128). For comparison reason, we only

classify the 'Happy' and 'Neutral' expression in Lifespan database. The recognition accuracy is 71.42%. The comparable result reported in Guo et al. (2013) for Lifespan database is 64.04%, and this result is obtained using the features in the same domain as the training.

The promising result indicates that, the hierarchical feature set gains itself by providing multi-level, multi-scale representation for the face image. The rich features are genuinely encoding the expression components rather than the environmental factors. Such an impressive property enables the proposed ADN framework successfully apply in transfer learning and multi-domain tasks.

4.4.6 Discussion

Upon now, we have comprehensively investigated the $L_{1/2}$ norm regularized ADN and its performance in facial expression recognition. The promising results arise more attentions to the principles behind the unsupervised deep structure.

Lots of research papers have formally analyze the behavior of deep network in learning hierarchical feature from unlabeled data. These algorithms are believed to be a mimic to the organization of the cortex Lee et al. (2008). The information of the learnt feature passed through the shallow layer to deep should be analog to the computations performed in visual areas of human brain.

To date, researchers have revealed certain properties of visual area V1 and V2. As demonstrated in Lee et al. (2008), on the shallow layer, localized, oriented, edge filters are extracted to model V1 cell receptive fields. Further on deeper layer, the network gains capability to pick up both contour and corners as well as junctions information.

Although we cannot clearly define the facial expression with such components, it indeed enriches the description of expression. Most of all, these features are spontaneously generated from deep network with sparse constraint without any other priors. It makes the unsupervised learning realistic in representation learning.

Look close to Fig. 4.8, the proposed deep network generates the hierarchical features which comprise of the oriented bars, junctions and contours. The behavior of the network just obeys the mechanism of visual cortex in area V1 and V2. Combined with data prior which is heavy-tailed distribution in both of the input data and generated features, we have confidence to enforce the expression representation learning with $L_{1/2}$ norm regularization. From this perspective, the performance of the proposed deep network for unsupervised FER is theoretical advantageous.

Chapter 5

Facial Feature Parsing and Landmark Detection via Low-rank Matrix Decomposition

5.1 Related Work

In general, the facial parsing tasks are implemented in two ways: landmark-based approach and segmentation-based algorithm Smith et al. (2013). On one side, the landmark-based method computes to mark each pixel on the face with semantic part label based on the pixel attribute. The performance heavily replies on well-defined initialization. The improved version of such approach includes the Markov Shape Model Liang et al. (2006) which considers the local line segments and appearance to alleviate the dependency of the initialization. However, the problem still remains since such technique asks for expensive computation cost and make it fails in multiple real-world applications. On the other side, the segmentation-based approach is proposed by considering the computational efficiency and robustness to pose, expression variations. In literature Liang et al. (2008), a component-based discriminant search algorithm is designed. Multiple facial component detectors are combined to detect

the facial parts. In the recent research, since facial parts have unique geometric configurations and appearances, the sparse matching and deep learning network are adopted to detect and correlate the facial parts and their label belongings. However, there are still drawbacks in these methods. First, facial features are hard to be given a uniform model. For example, there is no clear shape model to describe the mouth with different expressions. Pose variation may also cause appearance change in different scenarios. Once it failed to generate the model, it is impossible for matching process to locate the correct components on face. Second, it is difficult for a single detector to precisely locate all facial components. Thus, we either need to increase the number of detectors or get a failure parsing map in the process of matching.

Inspired by the saliency detection work in Shen and Wu (2012); Lang et al. (2012), we model the facial parsing as the salient feature detection on face. The motivation of this work compromises with the observation that, the facial features, e.g., eyes, nose and mouth have their unique appearances and thus making them visually salient comparing with the skin texture. We decompose the face image into small patches. It should be noted that, the facial features only occupy a small amount in these patches. It is coincide with the observation of sparsity. In the detailed process, multi-level features are explored to represent each patch. The feature vectors are then stacked to formulate the matrix. Both of the appearance information and spatial coherence of the sparsity would be reserved. Meanwhile, a linear feature transformation matrix is multiplied to boost the training in order to get a good feature representation with labeled data. After applying the low-rank matrix decomposition on the feature matrix, as the sparse noise, the facial feature components are expected to separate out from the recovered skin background. The whole algorithm is detection-based without any high level prior knowledge or models. That means if the facial components disappear due to occlusion, the algorithm cannot predict the positions of the features.

5.2 Parsing Algorithm

Before we introduce the parsing algorithm details, the overview of low-rank matrix decomposition is given in Section 5.2.1. After that, we describe the proposed feature representation and transformation matrix learning in two steps. To complete the parsing task, the refining process, as well as the landmark detection task are discussed in Section 5.2.4.

5.2.1 Matrix Decomposition by Low-rank Matrix Representation

Instead of using the spatial information of the image, the face is represented in feature space and denoted as **F**. We model the facial image as a combination of low-rank skin background **L** and facial features as sparse noise **S**. The prototype of matrix decomposition by low-rank representation (LLR) Liu et al. (2013) is formulated as solving the following problem,

$$\begin{aligned}(L^*, S^*) = \arg\min_{L,S}(rank(L) + \lambda\|S\|_0) \\ s.t. \quad F = L + S\end{aligned} \tag{5.1}$$

where, L^* and S^* are optimized results of skin background and the noise residual respectively. $\|\cdot\|_0$ represents the $L-0$ norm of the vector.

Solving Eq. (5.1) is NP-hard. A convex surrogate of the equivalent problem resulted in

$$\begin{aligned}(L^*, S^*) = \arg\min_{L,S}(\|L\|_* + \lambda\|S\|_1) \\ s.t. \quad F = L + S\end{aligned} \tag{5.2}$$

where $\|\cdot\|_*$ represents the nuclear norm and $\|\cdot\|$ indicates the $L-1$ norm. Solvers of LRR problem are proposed in many research articles. We adopt the most popular RobustPCA method Liu et al. (2013) which is more extendable and flexible in many cases. We are interested in the decomposition result S^* which contains the extracted

parsing map in gray-scale. The highlight pixels in S^* demonstrate the segmented facial components.

5.2.2 Facial Image Representation

Given a face image, we firstly apply Cascade Face Detector Viola and Jones (2004) to locate the face region. The detected face region is augmented to the uniform size 256×256. To acquire a complete representation of the face, similar to Shen and Wu (2012), we extract multi-modality visual features of the face region which are,

- Color Features in HSV color space. The image hue, saturation and light value are used to represent the color information;
- Texture Steerable Pyramids Simoncelli and Freeman (1995). The steerable pyramid filter is adopted to extract the texture information. We use the filter responses in 4 orientations and 3 scales results in 12 pyramid maps;
- Gabor Wavelet Features. Gabor filter is also performed to explore more detailed texture features. We totally use Gabor wavelet filters in 6 orientations and 3 scales on the image, yielding to 18 wavelet feature maps.

All of these features are properly normalized to reduce the cross-data variations. They are vertically stacked to formulate the feature vector. In the image space, we equally divided the face region into 4×4 non-overlapping cells. For each cell, the mean value of feature vectors f_i is computed to represent the entire cell. By reshaping the feature vectors, the matrix representation of feature maps is generated in the form of $F = [f_1, f_2, ...f_N]$, $F \in R^{D \times N}$, where D is the feature dimension (33) and N is the number of cells (4096 in our case). Matrix F is the facial image representation.

5.2.3 Learning Process of Linear Transformation Matrix

The matrix of the facial skin background is naturally low-rank in face image. To boost its characteristic, a linear transformation matrix $\mathbf{T}$ is employed to learn in

the feature space. In the learning process, we use the dataset with hand-labeled parsing map. For each training image, the image feature representation is extracted according to Sec. 5.2.2. A diagonal matrix $M = diag(m_1, m_2, ..., m_N)$ is generated to indicate feature vectors belongings to the skin background or not with the value 1 or 0. With such configuration, the transformation matrix T is learnable by solving the optimization problem,

$$T^* = \arg\min_T (\frac{1}{K}\sum_{k=1}^{K} \|TF_kM_k\|_* - \gamma\|T\|_*) \quad s.t. \quad \|T\|_2 = c \tag{5.3}$$

where, K is the total number of training images and k is the image index. To avoid a trivial result, we add a constraint to keep the $L-2$ norm of T be a small constant value ($c = 2$ in our experiments). For the solver of Eq. (5.3), one can refer to Shen and Wu (2012). The role of indicate matrix M is that it zeros out all the vectors in the feature matrix if it is not skin background. Therefore, the transformation matrix T is indeed forcing TFM to learn the features from background and meanwhile keeping it in low-rank form. The learnt T^* is used for all the testing images for parsing. In the testing stage, we multiply T with facial image representation F and derive the parsing map by applying Eq. (5.2) on T^*F.

5.2.4 Post-process and Landmark Detection

With the parsing map, our algorithm can be easily extended to detect the landmark points on face. In our work, we consider five landmarks which are left eye center, right eye center, nose tip, left mouth corner and right mouth corner. This setting is coincide with Sun et al. (2013) and most commonly used for landmark detection.

The landmark points are extracted based on the paring map. Considering the face detection algorithm we used for locate the face region, we have the prior knowledge that face is generally centered at the face region. So we multiply a Gaussian map Judd

et al. (2009) to enhance the detected feature localization. Then the kNN clustering algorithm is used to separate feature components. To acquire the clear boundary for each feature component, ellipse matching Ellis et al. (1991) is applied for each cluster. Although the facial components may not perfectly fit in the ellipse shape, it would not hurt the landmark localization. The eye center points are equivalent to the centers of the eye clusters and so as the nose tip. For the mouth corners, they are boundary points achieved by mouth cluster.

5.3 Experiments

We conduct two experiments on two challenging datasets FACES Guo et al. (2013) and LFPW Belhumeur et al. (2013) to evaluate the effectiveness of the proposed algorithm. Both of the datasets come with manually pointed landmarks. Parsing maps can be generated though. We perform transformation matrix learning based on FACES with randomly selected 1000 images. More concrete, the FACES image contains 80 human labeled landmarks along the entire face. We select the related points to formulate bounded rectangles which covered eyes, nose and mouth regions, as shown in Fig. 5.1. Based on these rectangles, the indicate matrix M is established according to the discussion in Sec. 5.2.3. All the images are uniformly resized into 256×256 as the input.

5.3.1 Experiment I: Qualitative Performance of Face Parsing

We test the proposed parsing algorithm on FACES initially. FACES images are challenging in terms of the large range of ages and variation in expressions. The proposed algorithm is designed to be robust against the variations. Notice that, eyebrows are not considered in our indicate matrix M, so that they are not the expected components in parsing although they are visually salient on face. To demonstrate the effectiveness of the transformation matrix T. We also apply LLR

Figure 5.1: Hand-labeled points on FACES image and the generated bounding rectangles for training.

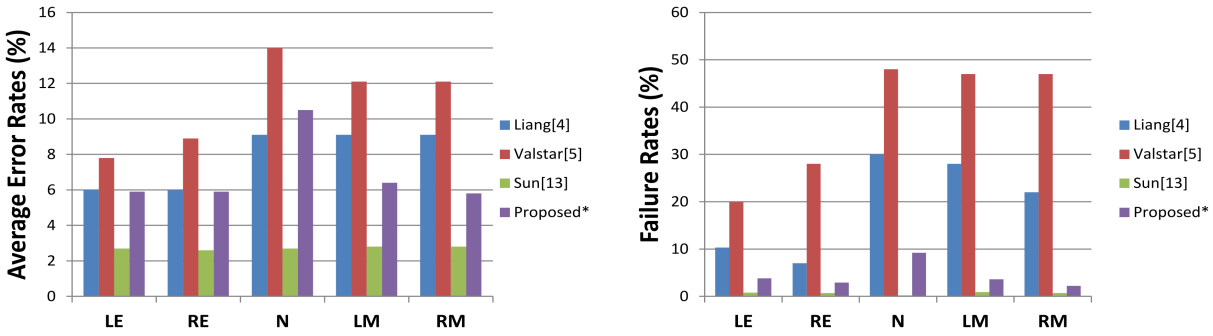


Figure 5.2: Qualitative comparison on LFPW dataset. Noticed that, our results received by testing on 500 non-occluded images.

to the facial feature representation matrix alone. The parsing results are shown in Fig. 5.3. Clearly, the matrix T enhances the capability to learn discriminant features which is beneficial to distinguish the facial components from the face region. For the LFPW dataset, it is more challenge because the faces contain more variations, such as pose, illumination change and occlusion. As we mentioned, the proposed algorithm is detection-based, once the facial components are occluded by other object and do not visually exist in face region, our method cannot predict their locations and segment them out. It is reasonable for real-world application scenarios. So we do not count it as failure. The parsing results are illustrated in Fig. 5.3.

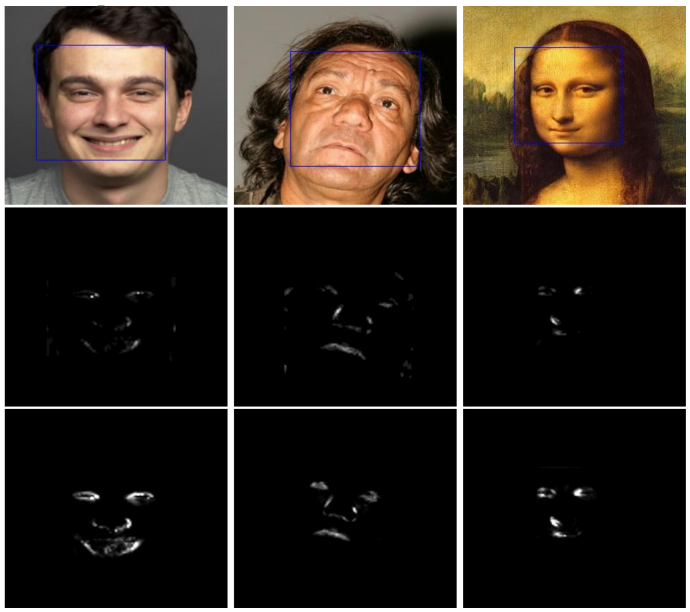


Figure 5.3: Parsing map demonstration. The first row contains the original input faces with Cascade Face Detector localized face regions. The second row contains the parsing map without transformation matrix T. The last row illustrates the parsing maps generated by the proposed algorithm. The parsing maps without T are polluted with unrelated pixels and the proposed method detects more regions on the facial components.

5.3.2 Experiment II: Quantitative Performance of Landmark Detection

The five-points landmark detection is validated by applying the method in Sec. 5.2.4 from derived parsing maps. The detected landmarks are demonstrated in Fig. 5.4. Quantitative performance is measured with average detection error which is defined as,

$$err = \sqrt{(x - x')^2 + (y - y')^2}/l \tag{5.4}$$

where (x, y) is the ground truth and (x', y') is the detected location. l represents the width of the bounding box. It is the same criteria as in Sun et al. (2013). If an error is larger than 5%, it is counted as a failed detection. The failure rate is also reported.

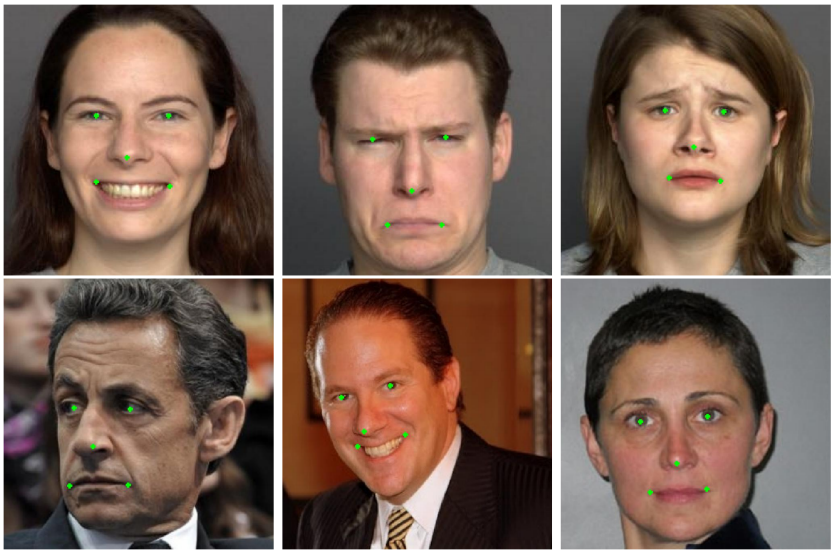


Figure 5.4: Landmark detection demonstration. The top row contains the images from FACES and the bottom row images come from LFPW.

For FACES, except the 1000 training images, the rest images (1052) in dataset are used for testing. The results are recorded in Table 5.1. For LFPW, since its nature that is not proper for our algorithm, we select 500 non-occluded images as testing samples. We give '*' mark on the comparison results with other methods. The comparable results are list in Fig. 5.2.

Table 5.1: Testing results on FACES dataset

	LE	**RE**	**N**	**LM**	**RM**
Average Errors (%)	3.2	4.4	5.7	1.9	2.1
Failure Rates (%)	1.7	1.5	7.8	2.1	1.8

From the experiments, the proposed algorithm receives favorable results on benchmark database. The effectiveness is completely validated.

5.4 Conclusion

In this chapter, we proposed a novel face parsing algorithm where the facial features are segmented out by modeling them as sparse noises via low-rank matrix decomposition. For precise parsing, a learnt linear transformation matrix T is added

to boost the performance. With the generated parsing maps, we further extended the work to accomplish the landmark detection on face. The proposed algorithm is tested on two benchmark datasets and demonstrated competitive performance comparing with the state-of-the-art techniques.

Chapter 6

Deep Tree-structured Face: A Unified Representation for Facial Biometrics

6.1 Introduction

Faces possess multiple channels information. It conveys biometric information such as human identity, age, expression, gender, etc. The ability to automatically understand face information is essential for computer vision, biometrics and psychology researches. Although the face analysis has been extensively studied for decades, face recognition, facial expression recognition and age estimation are still addressed separately in independent tasks. The critical problem of uniformly representing human face with multiple semantic meanings has not been well studied due to the highly coupled relationships of the latent factors in facial images.

Previous methods of facial image analysis allow the discriminant feature learning for specific facial biometric. Feature descriptors are proposed to model the face in shape Gökberk et al. (2006); Le et al. (2011), landmark Burgos-Artizzu et al. (2013), local texture information Guo et al. (2009); Ahonen et al. (2004) or even facial parts

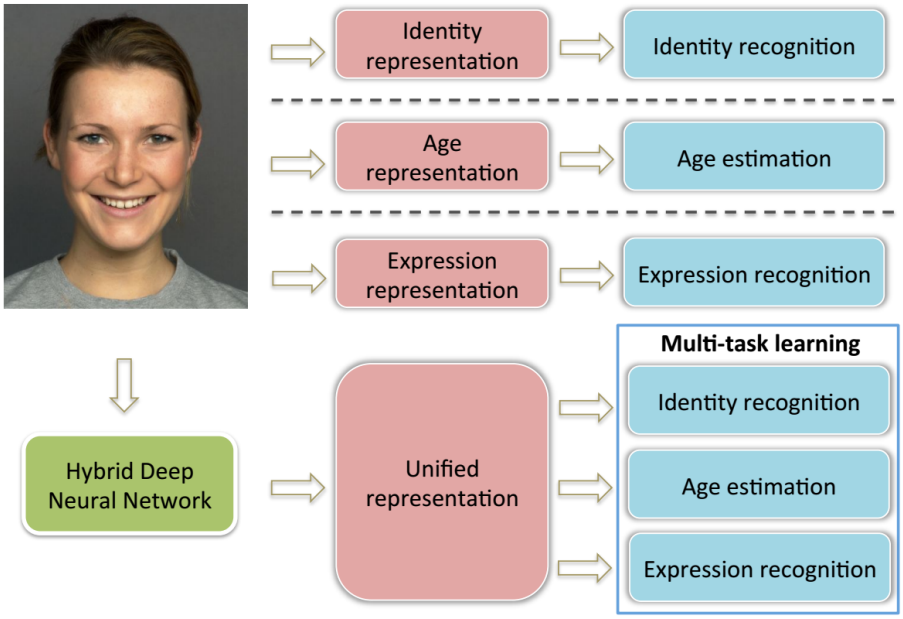


Figure 6.1: Motivation of this work. Traditional facial image analysis treats the face recognition, expression recognition and age estimation separatively. We propose to jointly learn a unified representation for the face and use it in multi-task biometrics.

with geometric relationship Senior (1999); Li et al. (2010). However, all these features are impractical to represent multi-factor facial semantics.

In the recent years, deep learning neural network receives a great deal of research interests. The deep models such as ConvNets LeCun et al. (1998), AlexNet Krizhevsky et al. (2012) and GoogLeNet Szegedy et al. (2015) have been proved effective to extract hierarchical visual features and successfully applied in facial biometrics. The success of deep learning comes from its strong discriminant learning capability and hierarchical representation for different patterns. However, in its layer-wised learning strategy, the feature map shrinkage by subsampling or low-dimension projection at the same spatial plane destroys the composite symbolic structures of input data, such as trees and graphs. It limits the generative representation for such data.

Motivated by the multi-task learning Collobert and Weston (2008) and deep learning concepts, we propose to build a new architecture which enables to jointly

learn a unified tree-structured face representation Socher et al. (2012). The proposed representation is constructed from the shallow neural network model which gains the advantage of utilizing over-complete low-level features Sun et al. (2014). It hierarchically combines local patches to generate the root node representation in tree structure. We recursively apply the semi-supervised AutoEncoder to enhance the semantic learning. The final learnt feature is claimed to uniformly represent the face image and be applicable to multi-task biometric recognitions. The motivation of the work is illustrated in Fig. 6.1.

6.2 Learning Tree-structured Face Representation

In this section, we describe the algorithm that is designed to formulate the tree-structured face presentation via the hybrid deep neural network. We firstly learn CNN filter kernels by clustering random patches from facial images. Images will be fed into the single layer CNN network, resulting in the translation-invariant low level multi-layer representation. Semi-supervised AutoEncoder is applied recursively to compose the tree-structured features that we would adopt for classification tasks.

6.2.1 Single-layer CNN Network Learning

The recent researches on 'Width versus Depth' in deep neural network reveal the important role of large number of nodes in the intermediate layers Coates et al. (2011); Socher et al. (2012). We follow the procedure proposed by Coates et al. Coates et al. (2011) to generate the low-level features with single layer CNN network. The algorithm contains K-means clustering to compute the filter kernels and single-layer CNN feature extraction with such learnt kernels.

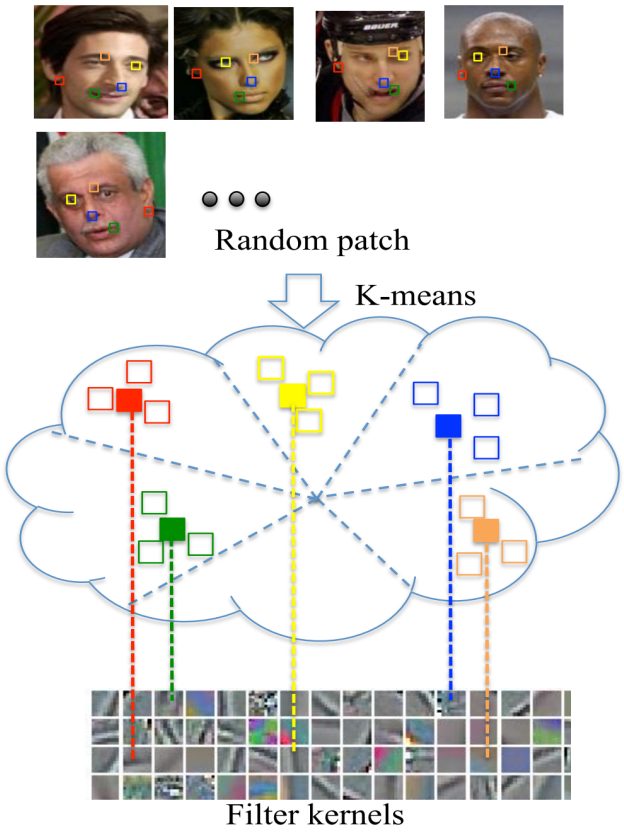


Figure 6.2: Unsupervised CNN Filter Kernel Learning. The solid squares represent centroids of clusters.

Unsupervised CNN Filter Kernel Generation

We randomly extract patches from given images. Normalization and whitening are used on these patches to reduce the variations. Then K-means clustering is applied over the patch set. Each cluster centroid is treated as the learnt filter kernel. Since we consider RGB-color images, both of the edge and color information are captured. We demonstrate filter kernel generation process in Fig. 6.2, and the learnt kernels are illustrated in Fig. 6.3. Clearly, the Gabor-like edge kernels with different scales and orientations are preserved. One particular result when using a large k value in K-means is that, the QR code-like kernels emerge in the learnt filter set, which have been proved to be deeper layer features in deep belief net Hinton et al. (2006). The unsupervised learnt kernels will be used in the following CNN network.

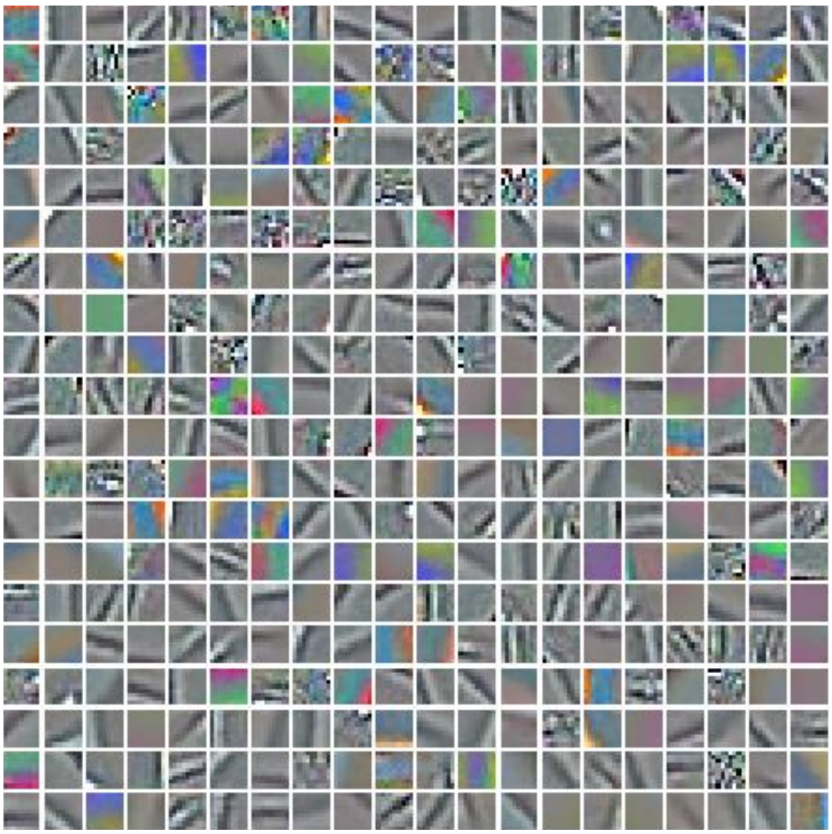


Figure 6.3: K-means learnt CNN filter kernels. $k = 400$, kernel size 9×9. Noticed that, both of the Gabor-like kernels and QR code-like kernels emerge which are similar to the deep belief net first two layers kernels.

Single-layer CNN Network

To generate the low-level features for the tree-structured face representation, we use the single-layer CNN network. The main idea is to explore the translation-invariant attributed features and the rich texture information from facial images. For each image of the size S_I (equally in height and width), we convolve it with squared filter kernels of the size S_k, which resulting in L convolution response of the dimensionality $S_I - S_k + 1$. We adopt max-pooling strategy after the convolution process. Each pooling is taken on a squared region of size S_p with a stride s. We then obtain a pooled response with equal height and width, sized as $r = (S_I - S_p)/s + 1$. The final output after the single-layer CNN network would be a 3D matrix of the dimensionality $L \times r \times r$. The 3D matrix will be served as the input for the subsequent neural network.

6.2.2 Tree-structured Face Representation via Semi-supervised AutoEncoder

Given a facial image, we are exploiting to find a hierarchical representation considering latent variables such as identity, expression and aging factors embedded in the image. The modeling process is implemented in a recursive manner via the semi-supervised AutoEncoder Socher et al. (2011). We firstly introduce the AutoEncoder network and proceed it to a semi-supervised version. Based on such a configuration, we describe how to apply it recursively to formulate the tree-structured representation for the face. We use the same procedure to all the facial images to generate the features for subsequent multi-task classification.

AutoEncoder Neural Network

AutoEncoder is a diabolo-shaped neural network. It is usually used to learn a reduced dimension representation for its input vector. Considering the goal of this research is to find a compact representation for face image, we follow the concept proposed by

A.Lemme et al. Lemme et al. (2012) to model the AutoEncoder with sparse and non-negative constraints which are efficient in encoding. Assume we are given a pair of input vectors $[x_p^k, x_q^k]$ from dataset with the length k, the AutoEncoder is used to find a new vector representation y^k for the input pair. It tries to learn the parameters in the network that reconstruct input vector catenation by minimizing the error function, denoted as

$$E_{rec}([x_p^k, x_q^k]) = \|g \circ f([x_p^k, x_q^k]) - [x_p^k, x_q^k]\|^2 \tag{6.1}$$

where $y^k = f(W[x_p^k, x_q^k] + b)$ is called encoder, $W \in R^{k \times 2k}$ is the tied weight matrix connects between input nodes and hidden nodes. $g(y^k) = W^T y^k$ acts as a decoder. f is usually a component-wise activation function deployed on the output layer. To enhance the compact feature learning, the sparsity constraint is incorporated by adopting the augmented logistic function as the activation which is in the form of

$$h([x_p^k, x_q^k]) = \frac{1}{1 + \exp^{-(a_i g_i - b_i)}} \tag{6.2}$$

The information transformation between the neurons are controlled by adjusting the parameter a_i and b_i. As the sparsity measurement, the mean activity level μ appears in the learning process of a_i and b_i. As proposed in the intrinsic plasticity mechanism, we acquire the sparsity by setting μ to $[0, 1]$ ranged small value. The updating rule for a_i and b_i is simply the gradient descent with learning rate η_{IP},

$$\triangle b_i = \eta_{IP}(1 - (2 + \frac{1}{\mu})h_i + \frac{1}{\mu}h_i^2) \tag{6.3}$$

$$\triangle a_i = \eta_{IP}\frac{1}{a_i} + g_i \triangle b_i \tag{6.4}$$

where h_i is the activation of the ith neuron. The lifetime sparsity is accomplished in gradient learning. We set $\eta_{IP} = 0.001$ and sparsity $\mu = 0.5$ for all the configurations.

To enhance the non-negative weight learning, we simply deploy the online error correlation rule in the weight matrix updating,

$$\triangle w_{ij} = \eta(x_i - \hat{x}_i)h_j + |\tilde{w}_{ij}| \tag{6.5}$$

where $|\tilde{w}_{ij}|$ converts negative value into positive. Learning rate η is set to 0.002. Such a configured AutoEncoder acquires encoding efficiency in a self-adaptive way that accelerates the procedure with sparse and non-negative properties Guo et al. (2015).

Semi-supervised AutoEncoder Neural Network

Upon now, the AutoEncoder neural network was completely under the unsupervised regime and totally induced to represent the general information towards capturing facial identity semantic in terms of reducing the facial image reconstruction error. Apparently, to find a unified facial representation which is suitable for different biometrics, such as expression recognition and age prediction, we need to disentangle such latent factors with more discriminative learning.

In the previous section, AutoEncoder is designed to model the input vector distribution with a compact representation. We can leverage the encoding process by adding on two softmax layers to predict expression and age distributions respectively,

$$d^{exp}(y; \theta_{exp}) = softmax(W_{exp}^{label} \cdot y) \tag{6.6}$$

$$d^{age}(y; \theta_{age}) = softmax(W_{age}^{label} \cdot y) \tag{6.7}$$

Assuming we have K_{exp} labels for expression prediction, d^{exp} would be a K_{exp}-dimensional multinomial distribution. Let t_k^{exp} be the k_{th} element of the ground truth label information for expression, the cross-entropy error is defined as

$$E_{cExp}(y, t^{exp}; \theta_{exp}) = -\sum_{k=1}^{K_{exp}} t_k^{exp} \log d_k^{exp}(y, \theta_{exp}) \tag{6.8}$$

The cross-entropy error for age prediction is defined in the same way, as

$$E_{cAge}(y, t^{age}; \theta_{age}) = -\sum_{k=1}^{K_{age}} t_k^{age} \log d_k^{age}(y, \theta_{age}). \tag{6.9}$$

The final semi-supervised AutoEncoder objective function is updated by adding these two cross-entropy terms together and applied over the entire training set

$$\begin{aligned} E([x_p, x_q], t^{exp}, t^{age}, \theta) = \lambda_1 E_{rec}([x_p, x_q]) \\ + \lambda_2 E_{cExp}(y, t^{exp}, \theta) \\ + \lambda_3 E_{cAge}(y, t^{age}, \theta). \end{aligned} \tag{6.10}$$

The hyperparameters λ_1, λ_2 and λ_3 weight and normalize the strengthes from different error terms with the constraint $\sum \lambda_i = 1$. In the training stage, the errors backpropagate and force the neural network to learn a new representation y which carries different semantics in terms of minimizing the objective function iteratively.

In practice, we use the same gradient descent-like rule to learn all the parameters, as described in Section 6.2.2. The semi-supervised AutoEncoder neural network is demonstrated in Fig. 6.4.

Tree-structured Representation Learning

AutoEncoder is easily to be applied recursively to learn a tree-structured representation once the tree's setting is available as a prior. It has already been used in the symbolic data representation learning, such as in natural language processing, it is used to predict sentence sentiment distribution Socher et al. (2011). Inspired by these work, we design a new setting which does not rely on the given tree structure, instead, the model can autonomously learn it from the data.

In Section 6.2.1 we have obtained the 3D matrix representation in dimensionality of $L \times r \times r$ to represent the facial image. In the image plane, the facial image is consist of $r \times r$ super-pixels. The computation loop of the tree-structured representation starts

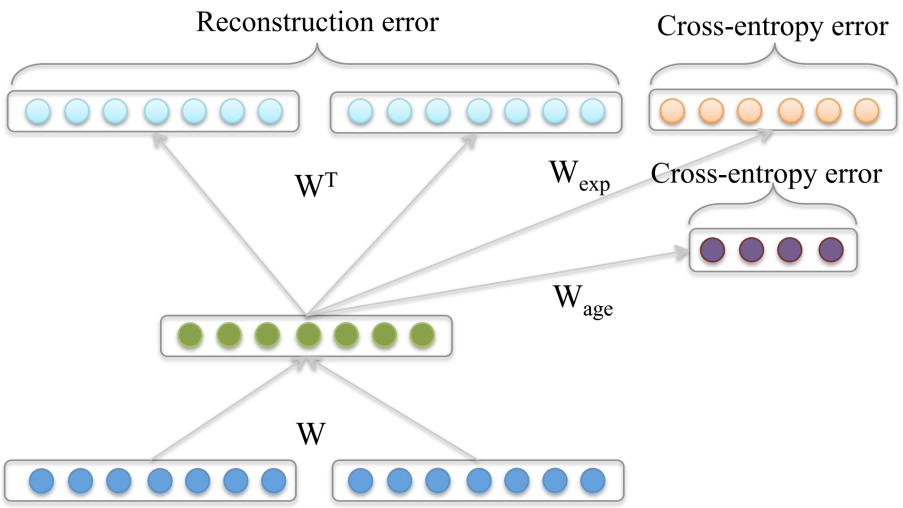


Figure 6.4: The structure of the semi-supervised AutoEncoder. We incorporate labeling information in terms of cross-entropy errors to enforce discriminant feature learning.

applying the semi-supervised AutoEncoder recursively on each pair of neighboring super-pixels, and recording the resulting errors. We compare and pick the super-pixel pair which has the smallest reconstruction error as the leaf nodes to generate a parent super-pixel. Then we shift the AutoEncoder to the parent super-pixel position, calculate the reconstruction errors between it and its neighbors, and update the error recording list. The computation loop continues to find the pair with smallest error and combine them until there is only one super-pixel remained, as shown in Fig. 6.5. The entire search and combine path is also recorded. The resulting super-pixel is the facial image representation and the final path is the learnt tree-structure. We would use them in the testing stage.

6.3 Experiments

In this section, we firstly introduce the multi-factor facial image dataset FACES Ebner et al. (2010) and the standard face recognition, expression recognition and age estimation experiments based on it. We then conduct extended experiments to analyze the performance of the unified representation. To receive the objective

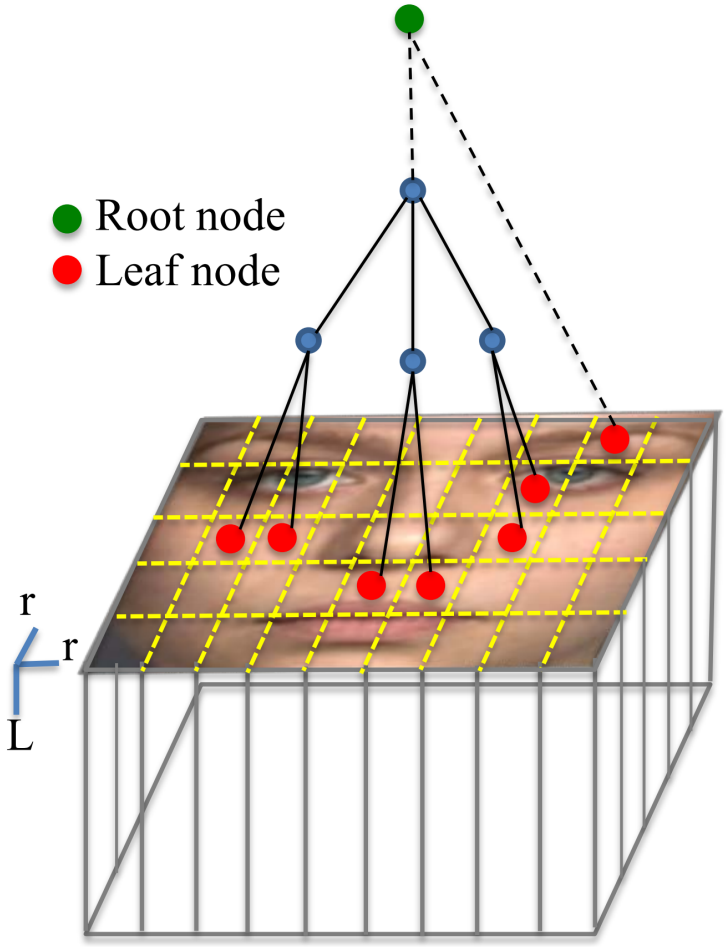


Figure 6.5: The computation model demonstration. The super-pixels are recursively combined to generate a tree-structured representation for the face image. Semi-supervised AutoEncoder is applied on each triplets to combine two super-pixels into one parent super-pixel.

Figure 6.6: FACES databases illustration. It contains the six expressions from Angry, Disgust, Fear, Happy, Neutral and Sad. The two individuals represent persons from different aging group.

comparison, we also tune the important parameters of the model and compare the proposed algorithm with other state-of-the-art techniques.

6.3.1 FACES Dataset

It is merely to see a facial image dataset considering multiple biometric factors including face identity, facial expression and aging influence in computer vision society. We turn to psychology study and find the FACES dataset carries human labeled ground truth for expression and age in a certain amount of images. It is the ideal testing benchmark for our research. The FACES dataset contains 171 individuals in the aging range from 18 to 94. Each individual performs 6 fundamental expressions (angry, disgust, happy, fear, neutral and sad) twice resulting totally 2052 images. All the images are captured under the laboratory environment in front view. The statistics of the dataset is list in Table. 6.1. The sample images are demonstrated in Fig. 6.6. We crop and resize all the images into size of 128×128.

Table 6.1: The statistics of the FACES dataset

Exp.	**Age Group**				**Total**
No.	**18-29**	**30-49**	**50-69**	**70-94**	**Images**
6	**51**	**35**	**31**	**54**	**2052**

Based on the dataset, we categorize the ages into four groups which are $18-29$, $30-49$, $50-69$ and $70-94$. It covers the typical aging range in demographic study. The age labels are binarized as $(0,0,0,1)$, $(0,0,1,0)$, $(0,1,0,0)$ and $(1,0,0,0)$ respectively. The same labeling method is used for expression categorization. Each of the expression label is a 6-digit binary code.

6.3.2 The Standard Tree-structured Representation Learning

The single-layer CNN feature extraction is conducted completely in unsupervised manner. We use LFW face dataset **?** to generate the CNN filter kernels aiming to explore more face related features. To acquire the rotation-invariant feature property, we adopt 'bootstrap' idea which spontaneously rotate each image by $\pm 15°$, $\pm 30°$ and $\pm 45°$. Thus, the LFW dataset is augmented by 6 times amount of images with different rotations. We then randomly select 500000 patches from the entire dataset. Each patch is of size 9×9. We run K-means over the patch set with $k = 400$ in the first experiment. The resulting 400 clustering centroids are used as filter kernels. The pooling kernel is chosen as 15×15 and the stride $s = 7$. After the single-layer CNN, we have the 3D matrix representation in the dimensionality of $400 \times 16 \times 16$ for each image.

Each semi-supervised AutoEncoder is applied spatially on the image plane through the tree-structured representation learning. It leads to a final root feature which has 400 dimensions. We claim this feature as the unified representation for the facial image. The face recognition, expression recognition and age estimation are implemented based on the extracted feature set.

We use entire one set for training which contains 1026 images, and the other set for testing. The RBF kerneled SVM classifier is chosen for face recognition. The root layer trained softmax classifiers are used for expression recognition and age estimation.

Table 6.2: Multi-task biometrics accuracies and average ranking

Methods	Face Rec.	Exp. Rec.	Age Est.	Ave. Rank	Feat. Dim.
Gabor	**73.2%**	**69.4%**	**49.2%**	**3**	**558**
LBP-PCA	**81.0%**	**63.3%**	**41.1%**	**3.33**	**400**
CNN	**92.3%**	**51.8%**	**54.5%**	**2.33**	**1000**
Proposed	**86.5%**	**76.5%**	**71.3%**	**1.33**	**400**

We compare our model to related models in the literatures. We choose hand-crafted Gabor feature Guo et al. (2013) as the baseline. LBP Shan et al. (2009) is also investigated as the unsupervised comparison. Finally, the state-of-the-art CNN based AlexNet Krizhevsky et al. (2012) (input size: 256×256) is also implemented in our multi-task biometrics. The classifiers used for comparison methods are all RBF kerneled SVMs. Table. 6.2 lists the main accuracy numbers. To receive a fair comparison, we also adopt average ranking as the metric in this multi-task experiment.

In the average ranking comparison, our proposed algorithm outperform all the methods. In the single task, the CNN feature performs 5.8% better than ours in face recognition, but badly performs in expression recognition. It can be explained that CNN has a strong discriminant learning capability to model the face image. But when the identity and expression factors coupled together, the expression related latent feature would be concealed by the greedy pooling strategy in CNN, which resulted in the failure to disentangle the semantic information.

6.3.3 Expression Recognition and Age Estimation Without Identity

There is a strong evidence that identity and expression are two deeply coupled factors in facial biometrics Rifai et al. (2012). We have already shown the disentangling capability of the proposed tree-structured representation in previous experiment. To make it complete, we conduct a new experiment that split the dataset into training and testing parts without any identity overlaps. Then the recognition tasks only focus

Table 6.3: Multi-task biometrics accuracies without identity.

Methods	Expression Recognition	Age Estimation	Feature Dimension
Gabor	**82.9%**	**70.4%**	**558**
LBP-PCA	**81.7%**	**65.3%**	**400**
CNN	**76.2%**	**69.8%**	**1000**
Proposed	**81.4%**	**75.2%**	**400**

on the feature itself and its performance to modeling subtle changes in expression and aging related activities. We randomly select 84 individuals from FACES dataset and use their facial images as training samples. Then the remaining 87 individuals and their images only emerge in the testing set. All the other settings are kept the same as the previous experiment. The final accuracy numbers are reported in Table. 6.3. The baseline Gabor feature and LBP received much improvements which is not surprised. The original literatures reveal that the skin wrinkle is the dominant feature to describe expression and aging activities. Gabor wavelets and LBP are typical texture information descriptors used broadly in expression and age related facial modeling. Our proposed facial representation achieves comparable accuracies which approves its effectiveness in modeling face texture information as well as in tree structure facial modeling.

6.3.4 Key Parameters Tuning

The proposed algorithm contains several hyper-parameters, such as convolution kernel size and kernel number, pooling kernel size, etc. We pick these values empirically in the previous experiments. Systematically parameter tuning is necessary to improve the performance. In this section, we conduct experiments to tune the convolution kernel size and kernel number to qualitatively analyze their roles and influences on the recognition tasks.

We reset the experimental configuration to the standard tree-structured representation learning, but vary the convolution kernel number in 100, 225, 400, 625 and 900.

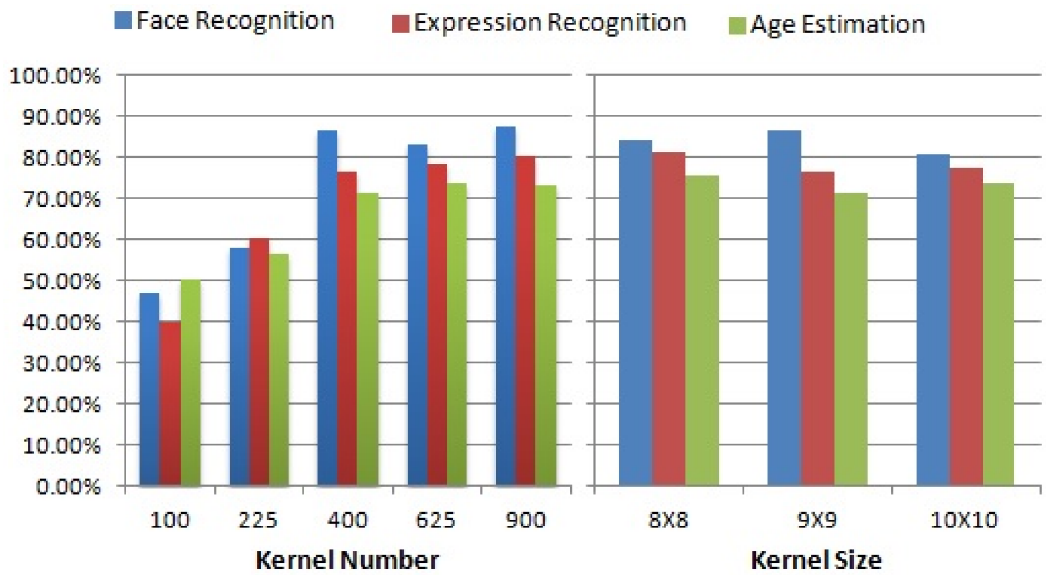


Figure 6.7: The recognition accuracies when tuning the key parameters.

The final multi-task classification accuracies are illustrated in the left side of Fig. 6.7. The accuracies reach stable values after using 400 kernels. One of the reasonable explanations can be addressed is that, 400 kernels over-completely represent the fundamental low-level features. Increasing kernel number more than 400 only resulted in the redundancy which decreases the discriminant feature learning capability.

The tuning on convolution kernel size is tested by using 8×8, 9×9 and 10×10 three configurations. We did not test it in extreme case since the deep learning research has already discussed about it Coates et al. (2011). The tuning of convolution kernel size affects all the following operations since the input dimension changed. To avoid the trivial configurations, after convolution operation, we linearly resize the 3D matrix to the same size as $400 \times 120 \times 120$. The experiment results are illustrated in the right side of Fig. 6.7. Clearly, the classification accuracies are not sensitive to this parameter.

6.4 Related Work

6.4.1 Facial Biometrics

To the best of our knowledge, this paper is the first one working on unified face representation for multi-task biometrics. The most related facial biometrics works consider the expression recognition task under aging influence and verse vase Guo et al. (2013); Guo and Wang (2012). In these studies, skin texture information are modeled to represent the face image since both of the aging and expression activities cause the skin wrinkles and distortion. To enforce the discriminative of the features, the correlation learning is deployed in the feature extraction process. Facial identity is treated as disturbance which is removed by splitting training and testing sets without identity overlap.

The newly proposed deep models mainly focus on extracting robust features to represent human face for the identification purpose Sun et al. (2014). In this scenario, the expression and aging influences will be treated as variations. Through the strong supervised learning, these factors are suppressed in feature learning.

Unsupervised deep network for expression recognition has been proposed in Rifai et al. (2012). In this paper, multi-scale contractive convolutional network and contractive discriminative analysis are combined to extract expression features against variations like identity, pose and face morphology.

6.4.2 Tree-structured Data Representation

The tree-structured face representation is initialized in this paper. The concept of such representation origins from natural language processing research. In Pollack's RAAMs model Pollack (1990), the recursive AutoEncoder was firstly introduced to learn vector representation with fixed data structure. Recently, Socher et al. Socher et al. (2011, 2012) proposed semi-supervised AutoEncoder to learn the tree-structured representation for sentence sentiment analysis. Later on, they continued the work on

recursive neural network with single-layer CNN for 3D object classification. Different from our algorithm, they adopted random weights in AutoEncoder with fixed tree structure. In our model, the weight and the tree structure are learnt from training data, which is more plausible to capture semantic contents from images.

6.5 Conclusion

We designed a new model based on a hybrid deep neural network to construct the tree-structured face representation. This architecture allows the learnt representation uniformly to be used in multi-task facial biometrics. The experiments were conducted comprehensively on a multi-factor dataset. Our proposed algorithm outperformed other state-of-the-art methods in terms of average rank evaluation. We also discussed the capability of the learnt representation in decoupling the latent information and the robustness when changed the key parameters. The final results shown its functional effectiveness in wide biometrics and computer vision applications.

One of the possible arguments may arise in facial image alignment requirement. The FACES dataset comes in the front view setting with slightly alignment of nose tip. The proposed tree-structured representation is learnt from training data. Different settings in pose, scales and occlusion definitely affect the tree structure or even make it failed to represent the face. Considering the single-layer CNN feature learning, the feature maps shrink at pooling stage which enables the tolerance to rotation and shift in a certain degree. Benefit from recording generated tree path, these deficiencies can also be handled by content-aware facial landmark detection or parsing. It is still a valuable novel face representation model in facial biometric research.

Chapter 7

Conclusion

A series of novel methods have been proposed to tackle the challenging task of image analysis with emphasization on facial biometric analysis. The image analysis problem can be decomposed into several different sub-problems. We focused the research in this work on three highly correlated sub-problems namely object detection, recognition, and detected facial image analysis. The goal was to design object models such that learning and inference can be performed efficiently for a large number of categories.

In the detailed implementation, we adopted saliency based object detection technique. The proposed approach conducted both of the local contrast technique based on bio-inspired attention feature and the global color distribution constraint. It was also observed that the proposed approach fully satisfies the criteria of biological observation on human vision and related application requirements. The proposed saliency based object recognition approach does not rely on any prior knowledge about the data distribution. Both of the texture, color and shape saliency can be efficiently pop out. The testing experiments were carried on with different benchmark databases of multiple object categories and validated with quantitative and qualitative analysis. The general performance reached the state-of-the-art in current unsupervised saliency

detection techniques. As the initial step to detect interesting objects, the proposed SMAP satisfied to provide stable detection results.

In the recognition stage, we investigated a novel approach based on the unsupervised deep ADN architecture. To strengthen the discriminant learning capability, we introduced the $L_{1/2}$ norm regularization to the prototype ADN. By conducting comprehensive evaluations on it, the proposed ADN structure was shown to be efficient in exploiting related features in an unsupervised manner. With the hierarchical features, our designed system receives competitive recognition accuracies compared to the previous state-of-the-art approaches on the saliency detection results. We also evaluated the method with a case study on the facial expression recognition task. The testing on benchmark databases received competitive performance and transfer learning ability. With the recognition process, the system gained the contextual awareness capability.

Once the human face was detected, we did facial feature parsing and landmark detection based on the detected face. We proposed a novel face parsing algorithm where the facial features are segmented out by modeling them as sparse noises via low-rank matrix decomposition. For precise parsing, a learned linear transformation matrix T is added to boost the performance. With generated parsing maps, we further extended the work to accomplish the landmark detection on face. The proposed algorithm is tested on two benchmark datasets and demonstrated competitive performance comparing with the state-of-the-art techniques. The completion of the work provided the capability to analyze the facial related activity (expression change, facial animation).

To better reason on the facial images, we designed a new model based on a hybrid deep neural network to construct the tree-structured face representation. This architecture allows the learnt representation uniformly to be used in multi-task facial biometrics. The experiments were conducted comprehensively on a multi-factor dataset. The proposed algorithm outperformed other state-of-the-art methods in terms of average rank evaluation. We also discussed the capability of the learnt

representation in decoupling the latent information and the robustness when changed the key parameters. The final results shown its functional effectiveness in wide biometrics and computer vision applications.

The image analysis utilized methods following the pipeline of the object detection, recognition and contextual analysis (detected face) have been fully implemented with advanced techniques in aforementioned steps. The comprehensive evaluations on each step indicated its efficiency and effectiveness to complete the proposed tasks. Considering the individual module, it is easy to transplant into different computer vision applications.

Bibliography

Achanta, R., Estrada, F., Wils, P., and Süsstrunk, S. (2008). Salient region detection and segmentation. In *Computer Vision Systems*, pages 66–75. Springer. 35

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned salient region detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*, pages 1597–1604. IEEE. 11, 35, 39

Ahonen, T., Hadid, A., and Pietikäinen, M. (2004). Face recognition with local binary patterns. In *Computer vision-eccv 2004*, pages 469–481. Springer. 80

Ahonen, T., Hadid, A., and Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041. 46

Amberg, B., Blake, A., Fitzgibbon, A., Romdhani, S., and Vetter, T. (2007). Reconstructing high quality face-surfaces using model based stereo. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE. 5

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43. 17

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press. 17

Avidan, S. and Shamir, A. (2007). Seam carving for content-aware image resizing. In *ACM Transactions on graphics (TOG)*, volume 26, page 10. ACM. xv, 43, 44

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202. 51

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2930–2940. 75

Bengio, Y. (2009). Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127. 14, 17

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. 17, 50, 67

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828. 13, 17

Bettadapura, V. (2012). Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*. 62

Borji, A. and Itti, L. (2012). Exploiting local and global patch rarities for saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 478–485. IEEE. 27

Broadbent, D. E. (2013). *Perception and communication*. Elsevier. 10

Burgos-Artizzu, X., Perona, P., and Dollár, P. (2013). Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520. 80

Calinon, S., Guenter, F., and Billard, A. (2007). On learning, representing, and generalizing a task in a humanoid robot. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 37(2):286–298. 30

Chen, C. and Hsiao, C. (1997). Haar wavelet method for solving lumped and distributed-parameter systems. *IEE Proceedings-Control Theory and Applications*, 144(1):87–94. 46

Cheng, M., Mitra, N. J., Huang, X., Torr, P. H., and Hu, S. (2015). Global contrast based salient region detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(3):569–582. 3, 12, 35, 39

Coates, A., Ng, A. Y., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *International conference on artificial intelligence and statistics*, pages 215–223. 82, 95

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM. 81

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE. 10, 46

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. 24

Deutsch, J. A. and Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological review*, 70(1):80. 10

Donoho, D. L. (2006). Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306. 50

Ebner, N. C., Riediger, M., and Lindenberger, U. (2010). Facesa database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, 42(1):351–362. 58, 89

Ellis, T., Abbood, A., and Brillault, B. (1991). Ellipse detection and matching with uncertainty. In *BMVC91*, pages 136–144. Springer. 75

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE. 10

Gao, D., Mahadevan, V., and Vasconcelos, N. (2008). The discriminant center-surround hypothesis for bottom-up saliency. In *Advances in neural information processing systems*, pages 497–504. 21

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741. 17

Goferman, S., Zelnik-Manor, L., and Tal, A. (2012). Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1915–1926. 35

Gökberk, B., İrfanoğlu, M. O., and Akarun, L. (2006). 3d shape-based face representation and feature extraction for face recognition. *Image and Vision Computing*, 24(8):857–869. 80

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *Neural information processing*, pages 117–124. Springer. 58, 63

Grossberg, S. (1995). The attentive brain. *American Scientist*, 83(5):438–449. 2

Guo, G., Guo, R., and Li, X. (2013). Facial expression recognition influenced by human aging. *Affective Computing, IEEE Transactions on*, 4(3):291–298. 5, 61, 68, 75, 93, 96

Guo, G., Mu, G., Fu, Y., and Huang, T. S. (2009). Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 112–119. IEEE. 25, 80

Guo, G. and Wang, X. (2012). A study on human age estimation under facial expression changes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2547–2553. IEEE. 96

Guo, R. and Qi, H. (2013). Partially-sparse restricted boltzmann machine for background modeling and subtraction. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 209–214. IEEE. 50

Guo, R., Wang, W., and Qi, H. (2015). Hyperspectral image unmixing using autoencoder cascade. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2015 7th Workshop on*, pages 1–4. IEEE. 87

Harel, J., Koch, C., and Perona, P. (2006). Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552. 35

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800. 4, 17

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554. 83

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE. xiv, 11, 12, 22, 25, 35

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154. 10, 17

Itti, L. and Baldi, P. F. (2005). Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, pages 547–554. 11

Itti, L., Dhavale, N., and Pighin, F. (2004). Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Optical science and technology, SPIE's 48th annual meeting*, pages 64–78. International Society for Optics and Photonics. 11

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259. 11, 35

James, W. (2013). *The principles of psychology*. Read Books Ltd. 10

Jamieson, A. R., Drukker, K., and Giger, M. L. (2012). Breast image feature learning with adaptive deconvolutional networks. In *SPIE Medical Imaging*, pages 831506–831506. International Society for Optics and Photonics. 48

Jia, Y. and Darrell, T. (2011). Heavy-tailed distances for gradient based image descriptors. In *Advances in Neural Information Processing Systems*, pages 397–405. 52

Judd, T., Durand, F., and Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. 36

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE. xiv, 35, 36, 37, 74

Koch, C. and Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer. 11

Kootstra, G., Nederveen, A., and De Boer, B. (2008). Paying attention to symmetry. In *British Machine Vision Conference (BMVC2008)*, pages 1115–1125. The British Machine Vision Association and Society for Pattern Recognition. 11

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. 81, 93

Lang, C., Liu, G., Yu, J., and Yan, S. (2012). Saliency detection by multitask sparsity pursuit. *Image Processing, IEEE Transactions on*, 21(3):1327–1338. 71

Le, V., Tang, H., and Huang, T. S. (2011). Expression recognition from 3d dynamic faces using robust spatio-temporal shape features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 414–421. IEEE. 80

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. 4, 17, 81

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, 1:0. 14, 18

LeCun, Y., Kavukcuoglu, K., Farabet, C., et al. (2010). Convolutional networks and applications in vision. In *ISCAS*, pages 253–256. 17, 62

Lee, H., Ekanadham, C., and Ng, A. Y. (2008). Sparse deep belief net model for visual area v2. In *Advances in neural information processing systems*, pages 873–880. 68

Lemme, A., Reinhart, R. F., and Steil, J. J. (2012). Online learning and generalization of parts-based image representations by non-negative sparse autoencoders. *Neural Networks*, 33:194–203. 86

Li, L.-J., Socher, R., and Fei-Fei, L. (2009). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2036–2043. IEEE. 1

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in cognitive sciences*, 6(1):9–16. 11, 21, 38

Li, Z., Imai, J.-i., and Kaneko, M. (2010). Robust face recognition using block-based bag of words. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1285–1288. IEEE. 81

Liang, L., Wen, F., Tang, X., and Xu, Y.-q. (2006). An integrated model for accurate shape alignment. In *Computer Vision–ECCV 2006*, pages 333–346. Springer. 70

Liang, L., Xiao, R., Wen, F., and Sun, J. (2008). Face alignment via component-based discriminative search. In *Computer Vision–ECCV 2008*, pages 72–85. Springer. 5, 70

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. (2013). Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):171–184. 5, 72

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., and Shum, H.-Y. (2011). Learning to detect a salient object. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):353–367. 30, 37

Ma, Y.-F. and Zhang, H.-J. (2003). Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381. ACM. 12, 35, 36, 37

Minear, M. and Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4):630–633. 58

Mirota, D., Taylor, R. H., Ishii, M., and Hager, G. D. (2009). Direct endoscopic video registration for sinus surgery. In *SPIE Medical Imaging*, pages 72612K–72612K. International Society for Optics and Photonics. 1

Myers, W. (1979). Interactive computer graphics: Flying high-part i. *Computer*, (7):8–17. 27

Newman, P., Cole, D., and Ho, K. (2006). Outdoor slam using visual appearance and laser ranging. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 1180–1187. IEEE. 1

Niebur, E. and Koch, C. (1998). Computational architectures for attention. *The attentive brain*, pages 163–186. 11

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987. 46

Oliva, A. and 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. 10

Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123. 11

Phung, S. L. and Bouzerdoum, A. (2009). Matlab library for convolutional neural networks. *University of Wollongong, Tech. Rep., URL: http://www. elec. uow. edu. au/staff/sphung*. xii, 63

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, 46(1):77–105. 96

Reynolds, J. H. and Desimone, R. (2003). Interacting roles of attention and visual salience in v4. *Neuron*, 37(5):853–863. 21

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025. 24

Rifai, S., Bengio, Y., Courville, A., Vincent, P., and Mirza, M. (2012). Disentangling factors of variation for facial expression recognition. In *Computer Vision–ECCV 2012*, pages 808–822. Springer. 13, 93, 96

Rutishauser, U., Walther, D., Koch, C., and Perona, P. (2004). 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*, volume 2, pages II–37. IEEE. 3

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464. 30, 31

Senior, A. W. (1999). Face and feature finding for a face recognition system. In *Proc. Int. Conf. Audio Video-based Biometric Person Authentication*, pages 22–23. 81

Serre, T., Wolf, L., and Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 994–1000. IEEE. 62

Shan, C., Gong, S., and McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816. xii, 63, 93

Shen, X. and Wu, Y. (2012). A unified approach to salient object detection via low rank matrix recovery. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 853–860. IEEE. 35, 71, 73, 74

Simoncelli, E. P. and Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *icip*, page 3444. IEEE. 73

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216. 11

Smith, B., Zhang, L., Brandt, J., Lin, Z., and Yang, J. (2013). Exemplar-based face parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3484–3491. 70

Socher, R., Huval, B., Bath, B., Manning, C. D., and Ng, A. Y. (2012). Convolutional-recursive deep learning for 3d object classification. In *Advances in Neural Information Processing Systems*, pages 665–673. 82, 96

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics. 85, 88, 96

Sun, Y., Wang, X., and Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483. 74, 77

Sun, Y., Wang, X., and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898. 82, 96

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9. 81

Tang, Y. (2013). Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*. 63

Thirde, D., Borg, M., Ferryman, J. M., Fusier, F., Valentin, V., Brémond, F., Thonnat, M., and Team, O. (2006). A real-time scene understanding system for airport apron monitoring. In *ICVS*, volume 6, page 26. Citeseer. 1

Treisman, A. and Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological review*, 95(1):15. 10

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136. 2

Valstar, M., Martinez, B., Binefa, X., and Pantic, M. (2010). Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2729–2736. IEEE. 5

Vezhnevets, V. and Konouchine, V. (2005). Growcut: Interactive multi-label nd image segmentation by cellular automata. In *proc. of Graphicon*, pages 150–156. Citeseer. 41

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154. 73

Wang, X. and Tang, X. (2009). Face photo-sketch synthesis and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):1955–1967. 5

Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in neural information processing systems*, pages 2080–2088. 20

Xu, Z., Chang, X., Xu, F., and Zhang, H. (2012). regularization: A thresholding representation theory and a fast solver. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(7):1013–1027. 50, 51, 52, 64

Yu, H. and Yang, J. (2001). A direct lda algorithm for high-dimensional datawith application to face recognition. *Pattern recognition*, 34(10):2067–2070. xii, 63

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pages 818–833. Springer. 13, 17, 18, 48, 66

Zeiler, M. D., Taylor, G. W., and Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2018–2025. IEEE. 4, 17, 18, 47, 48, 49, 60, 65, 66

Zhai, Y. and Shah, M. (2006). Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 815–824. ACM. 35

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., and Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32. 11, 21

Appendix

Publications

- Scene Understanding
  - **Rui Guo**, Hairong Qi. Saliency Detection Based on Self-similarity. EURASIP Journal on Image and Video Processing (under review)
  - **Rui Guo**, Hairong Qi. Facial Feature Parsing and Landmark Detection via Low-Rank Matrix Decomposition. In ICIP 2015
  - **Rui Guo**, Hairong Qi. Partially-sparse Restricted Boltzmann Machine for Background Modeling and Subtraction. In ICMLA 2013
- Facial Biometrics
  - **Rui Guo**, Liu Liu, Wei Wang, Ali Taalimi, Chi Zhang and Hairong Qi. Deep Tree-structured Face: A Unified Representation for Multi-task Facial Biometrics. in WACV 2016
  - **Rui Guo**, Hairong Qi. Exploring Facial Expression via Unsupervised Deep Network with L1/2 Norm Regularization. IEEE Trans. on Affective Computing (in submission)
  - X. Wang, **R. Guo** and C. Kambhamettu. Deeply-learned Features for Age Estimation. in WACV 2015
  - X. Wang, V. Ly, **R. Guo** and C. Kambhamettu. 2D-3D Face Recognition via Restricted Boltzmann Machines. in ICCVTA 2014

- G-D. Guo, **R. Guo** and X. Li. Facial Expression Recognition Influenced by Human Aging. IEEE Trans. Affective Computing, Vol. 4(3), pages 291-298, 2013.

- Hyperspectral Image Analysis
  - Ying Qu, **Rui Guo**, Wei Wang, Hairong Qi, Bulent Ayhan, Chiman Kwan and Steven Vance. Anomaly Detection In Hyperspectral Image Through Spectral Unmixing and Low Rank Decomposition. in IGARSS 2016
  - **Rui Guo**, Wei Wang and Hairong Qi. Hyperspectral Image Unmixing using Autoencoder Cascade. in WHISPERS 2015 **(Best Paper Award)**

- Pervasive Healthcare
  - Shuangjiang Li, **Rui Guo**, and etc. Demo: MoodMagician - A Pervasive and Unobtrusive Emotion Sensing System using Mobile Phones for Improving Human Mental Health. in Sensys Demo 2014
  - **Rui Guo**, Shuangjiang Li, Li He, Wei Gao, Hairong Qi and Gina Owens. Pervasive and Unobtrusive Emotion Sensing for Human Mental Health. in Pervasive Health 2013

- Others
  - Chi Zhang, Hao Zhang, **Rui Guo**, and Lynne E. Parker, A Unified Representation for Robot Learning of Action Labels and Motion Trajectories from Internet 3D Human Skeletal Data. in ROMAN 2016
  - Wei Wang, Ali Taalimi, Kuan Lu, **Rui Guo** and Hairong Qi. Learning Patch-Dependent Kernel Forest for Person Re-Identification. in WACV 2016
  - X. Li, **R. Guo**, C. Chen. Robust Pedestrian Tracking and Recognition from FLIR Video: A Unified Approach via Sparse Coding. Sensors, Vol.6, pages 11245-11259, June, 2014

Vita

Rui Guo was born in Baoji, Shaanxi Province, China. He received his B.S. and M.S. degree in Beihang University, Beijing, China (BUAA) in 2006, 2009, respectively. Since 2009, he worked at West Virginia University to start his Ph.D study under the supervision of Dr. Xin Li. In 2011, he transferred to University of Tennessee, Knoxville and continued pursuing his Ph.D at the Advanced Imaging and Collaborative Information Processing (AICIP) Lab under the supervision of Prof. Hairong Qi. His research interests include computer vision, image processing, pattern recognition and machine learning.