Semantic Sparse Learning in Images and Videos

by

Qiang Zhang

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved May 2014 by the
Graduate Supervisory Committee:

Baoxin Li, Chair
Pavan Turaga
Yalin Wang
Jieping Ye

ARIZONA STATE UNIVERSITY

August 2014

ABSTRACT

Many learning models have been proposed for various tasks in visual computing. Popular examples include hidden Markov models and support vector machines. Recently, sparse-representation-based learning methods have attracted a lot of attention in the computer vision field, largely because of their impressive performance in many applications. In the literature, many of such sparse learning methods focus on designing or application of some learning techniques for certain feature space without much explicit consideration on possible interaction between the underlying semantics of the visual data and the employed learning technique. Rich semantic information in most visual data, if properly incorporated into algorithm design, should help achieving improved performance while delivering intuitive interpretation of the algorithmic outcomes.

My study addresses the problem of how to explicitly consider the semantic information of the visual data in the sparse learning algorithms. In this work, we identify four problems which are of great importance and broad interest to the community. Specifically, a novel approach is proposed to incorporate label information to learn a dictionary which is not only reconstructive but also discriminative; considering the formation process of face images, a novel image decomposition approach for an ensemble of correlated images is proposed, where a subspace is built from the decomposition and applied to face recognition; based on the observation that, the foreground (or salient) objects are sparse in input domain and the background is sparse in frequency domain, a novel and efficient spatio-temporal saliency detection algorithm is proposed to identify the salient regions in video; and a novel hidden Markov model learning approach is proposed by utilizing a sparse set of pairwise comparisons among the data, which is easier to obtain and more meaningful, consistent than tradition labels, in many scenarios, e.g., evaluating motion skills in surgical simulations.

In those four problems, different types of semantic information are modeled and incorporated in designing sparse learning algorithms for the corresponding visual com-

puting tasks. Several real world applications are selected to demonstrate the effectiveness of the proposed methods, including, face recognition, spatio-temporal saliency detection, abnormality detection, spatio-temporal interest point detection, motion analysis and emotion recognition. In those applications, data of different modalities are involved, ranging from audio signal, image to video. Experiments on large scale real world data with comparisons to state-of-art methods confirm the proposed approaches deliver salient advantages, showing adding those semantic information dramatically improve the performances of the general sparse learning methods.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my adviser, Professor Baoxin Li for being a great adviser. As a scientist, his great passion on researches and hardworking has inspired me through the five years. As a professor and graduate adviser, he never hesitates spending efforts and sharing his experience and opinions to help me build my own views. I appreciate his inspiring encouragement to my Ph.D. life, his elaborate mentorship and constructive comments on my work, and his patience in helping me improve my professional skills from varying aspects. In addition to that, the conversations with him are very helpful for my future career and life. Thanks, Prof. Li, for all your support and help in the past five years!

I would like to thank my committee members, Professor Pavan Turaga, Professor Yalin Wang, and Professor Jieping Ye for serving as members of my dissertation committee and providing enlightening discussions and suggestions.

I would like to thank my colleagues at Arizona State University, as well as collaborators at Sharp Lab America, Qualcomm and Samsung. In particular, I would like to thank my previous and current group fellows, Xiaolong Zhang, Zheshen Wang, Nan Li, Naveen Kulkarni, Devi Archana Paladugu, Dr. Peng Zhang, Parag Chandakkar, Lin Chen, Hima Bindu Paladugu, Qiongjie Tian, Ragave Venkatesan, Yilin Wang and Xu Zhou, the close interactions between us really make our group like a family. I would also like to express my sincere thanks to Liang Sun, Jiayu Zhou and Ziming Zhao for their constructive discussions on my researches. I appreciate the great help from Dr Peter Van Beek, Dr. Xiquan Cui, Dr. ZhengPing Ji, Dr. Ilia Ovsiannikov, Dr. Ibrahim Sezan, Dr. Hae Jong Seo, Dr. Lilong Shi, Dr. Xinyu Xu, Dr. Chang Yuan, who made my summer 2011, summer 2012 and spring 2014 internship precious experiences.

My heartfelt appreciations go to my dear parents Xuedou Zhang and Qingying Chen. Their love and support throughout my life made my Ph.D. studies possible. I am also deeply grateful to Hui Lin (Iowa State University), Ruoyu Wu, Yadong Shen

and other fellow students and friends at ASU. Friendships with them made me feel ASU and Tempe as my hometown in US.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter 1

INTRODUCTION

Many learning models have been proposed for various tasks in visual computing. Popular examples include hidden Markov models and support vector machines. Recently, sparse-representation-based learning methods have attracted a lot of attention in the computer vision field, largely because of their impressive performance in many applications. For example, the so-called SRC algorithm Wright *et al.* (2009b) uses a simple sparse representation method for face recognition and was able to outperform many state-of-the-art methods when it was first proposed.

However, in the literature, many of such sparse learning methods focus on designing or application of some learning techniques for certain feature space without much explicit consideration on possible interaction between the underlying semantics of the visual data and the employed learning technique. For example, in the SRC algorithm, the physical conditions of the images are not explicitly considered. As a result, its performance degrades on smaller dictionaries. Also, the physical conditions of the images cannot be recovered. We believe that, rich semantic information in most visual data, if properly incorporated into algorithm design, should help achieving improved performance while delivering intuitive interpretation of the algorithmic outcomes.

My study addresses the problem of how to explicitly consider the semantic information of the visual data in the sparse learning algorithms. In this work, we identify four problems which are of great importance and broad interest to the community. In those four problems, different types of semantic information, from basic label of the training images to physical process of the image formation and to the high-level human visual system, are modeled and incorporated in designing the sparse learning algorithms for the corresponding visual computing tasks.

Our first semantic sparse learning method is related to dictionary learning. Dictionary learning for sparse representation has been widely used in computer vision, such as image denoising Elad and Aharon (2006), image inpainting Mairal *et al.* (2008c), image compression Bryt and Elad (2008a). However, the dictionary learned thereby is not optimized for a classification task. In other words, the learned dictionary may not have the best discriminative power despite its representational power. To this end, we demonstrate how can we combine the label of the data to improve the discriminative capability of the learned dictionary in the proposed method. We formulate this problem by combining the reconstruction error, classification error and sparsity constraint and solve it efficiently as a variant of K-SVD Aharon *et al.* (2005). The proposed method is evaluated on YaleB dataset and AR dataset with comparison to the state-of-arts, which demonstrates its effectiveness.

In our second sparse learning method, we explore the physical process of image formation and apply this process to improve the performances of face recognizer. Existing algorithms for face recognition either utilize the statistics of the human fixation of the images while ignoring the imaging process, or try to model the physical processes of image formation under different conditions with a large dictionary. Recognizing that non-sparse conditions such as illumination change and large occlusion are critical for face recognition, and that for a typical application we may assume only a finite number of such conditions (e.g., a relatively small number of illumination conditions or other conditions), we propose a model for representing a set of face images by decomposing them into three components: a common component shared by images of the same subject, a low-rank component capturing non-sparse global changes, and a sparse residual component. The learned common and low-rank components form a compact and discriminative representation of the original set of images. A classifier is then built based on the comparison of subspace spanned by these components and by a novel image to be classified. This is very compact compared with the number of atoms in an over-determined dictionary such as that in Wright *et al.* (2009b). Fur-

ther, by explicitly modeling non-sparse conditions, the proposed approach is able to handle both illumination changes and large occlusions, which would fail methods like Nagesh and Li (2009).

Sparse representation together with semantic information can also be applied to detect spatiotemporal saliency in video. In the recent years modeling and detection of visual saliency has attracted a lot of interest in the vision community, where a lot of different models have been proposed for computing visual saliency. Different sets of semantic information can be used for saliency detection, e.g., the saliency region of video is much smaller than the whole volume of the video (i.e., sparsity); the primary visual cortex (V1), where the saliency map for human vision exists, is orientation selective and lateral surround inhibition Simoncelli and Schwartz (1999). Based on those semantic information, as our third problem, we propose a novel spatiotemporal visual saliency detector for video analysis, based on *the phase information of the video*. Compared with existing methods in the literature, the proposed method is much more efficient, more capable of modeling complex dynamics and training-data-free.

For our final problem, we show that the semantic information as the relative ranking can be applied to learn hidden Markov model for modeling the motion skills. Existing methods for automatic evaluating motion skills typically require the skill labels for the training data. However, labeling the skill of the motion is currently done by human professionals, which is not only a costly practice but also one that is subjective and less quantifiable. Thus it is difficult, if not impossible, to obtain sufficient and consistent skill labels for a large amount of data for reliable training. Instead, we propose a novel formulation termed *Relative Hidden Markov Model* and develop an algorithm for obtaining a solution under this model. The proposed method utilizes only a sparse set of relative ranking (based on an attribute of interest, or motion skill in the surgical training application) between pairs of the inputs, which is easier to obtain and often more consistent. The proposed algorithm effectively learns a model from the training data so that the attribute under consideration (i.e., the

motion skill in our application) is linked to the likelihood of the inputs under the learned model.

Several real world applications are selected to demonstrate the effectiveness of the proposed methods, including, face recognition, spatio-temporal saliency detection, abnormality detection, spatio-temporal interest point detection, motion analysis and emotion recognition. In those applications, data of different modalities are involved, ranging from audio signal, image to video. Experiments on large scale real world data with comparisons to state-of-art methods confirmed the proposed approaches deliver manifest advantages, showing adding those semantic information dramatically improve the performances of the general sparse learning methods.

In the rest of the report, we will first introduce the basic of sparse learning in Chapter 2, then we will elaborate the above four sub-problems in details in Chapter 3, 4, 5 and 6. At the end of the dissertation, Chapter 7 presents a summary of current progress and a plan of future work along this direction.

In the presentation, we use upper case bold font for matrices, e.g., $\mathbf{X}$, lower case bold font for vectors, e.g., $\mathbf{x}$ and normal font for scalars, e.g., x, upper cases for constants, e.g., N.

Chapter 2

BASIC OF SPARSE LEARNING

In this chapter, we will introduce some basics of sparse learning. Sparse learning has a lot of applications in computer vision, e.g., compressed sensing Donoho (2006), face recognition Wright *et al.* (2009b), image denoising Elad and Aharon (2006), image compression Elad and Aharon (2006), image super resolution Yang *et al.* (2012), visual tracking Liu *et al.* (2011), image classification Yang *et al.* (2009), visual saliency Yan *et al.* (2010), action recognition Guha and Ward (2012) and so on. Sparse learning problems can be typically formulated as adding sparsity term to traditional learning methods,

$$\mathbf{x} : \min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}) \text{ s.t. } \mathbf{x} \in \Omega \tag{2.1}$$

where $\mathbf{x}$ is a vector, $f(\cdot)$ is the term from traditional learning methods and $g(\cdot)$ is the sparsity term. Different terms have been proposed for $g(\cdot)$, e.g., ℓ_0, which measures the number of nonzero elements Chen *et al.* (1991), i.e., $\|\mathbf{x}\|_0 = \sum_i (x_i \neq 0)$. However, ℓ_0 is not convex which typically makes the problem NP hard. Instead, ℓ_1 norm, which is the sum of absolute value of the elements, or $\|\mathbf{x}\|_1 = \sum_i |x_i|$, is more often used.

One of the simplest and most common examples is:

$$\mathbf{x} : \min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 \text{ s.t. } |\mathbf{x}\|_0 \leq \tau \tag{2.2}$$

This problem (nonconvex) can be efficiently solved via matching pursuit algorithm: in each iteration, a column of $\mathbf{D}$, $\mathbf{d}_i$, is selected which has largest correlation with $\mathbf{x}$, then $\mathbf{x} = \mathbf{x} - \frac{\mathbf{x}^T\mathbf{d}_i}{\|\mathbf{d}_i\|_2}$, and we repeat this procedure until τ columns are selected or $\|\mathbf{x}\|_2$ is smaller than some value.

If we replace ℓ_0 by ℓ_1 and apply Lagrange multipliers, we will get the following convex problem:

$$\mathbf{x} : \min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}x\|_2^2 + \gamma|\mathbf{x}\|_1 \tag{2.3}$$

which is known as least absolute shrinkage and selection operator (or LASSO). This problem can be solved by applying soft thresholding: $\mathbf{x} = S_\gamma((\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y})$, where $S_t(x) = \text{sign(x)}\max\{0, |x| - t\}$. Other algorithms have also been proposed for solving ℓ_1 norm related problems, e.g., GPSR Chen *et al.* (1991), ℓ_1 magic Candes and Romberg (2005) and so on.

Sparse learning method can also be generalized to matrix. Instead of measuring the number of nonzero elements or sum of absolute values of the elements, the rank of the matrix has drawn more interest, e.g, matrix completion Candes and Plan (2009), foreground segmentation in video Wright *et al.* (2009a), face recognition Zhang and Li (2012) and so on. The rank of matrix can be measured as the number of linear independent columns or rows of the matrix, whichever is smaller. It can also be measured as the number of nonzero singular values for the matrix Candes and Plan (2009), i.e., $\text{rank}(\mathbf{X}) = \|\Sigma\|_0$, where $\mathbf{X} \to \mathbf{U}\Sigma\mathbf{V}^T$ is the singular value decomposition. However, the rank measurement is not convex, thus a relaxed convex measurement: trace norm (or nuclear norm) is more often used Cai *et al.* (2008). The nuclear norm or trace norm of a matrix is defined as the sum of the singular value of the matrix, i.e., $\|\mathbf{X}\|_* = \|\Sigma\|_1$ (please note that for Σ, all of its elements are non-negative and its off-diagonal elements are zero.). The nuclear norm or trace norm related problems can be solved by iterative thresholding method Lin *et al.* (2009).

Sparse learning problem is also related to beta process in statistical learning. A draw from beta process $\text{BP}(a, b)$ can be represented as:

$$\begin{aligned}
\mathbf{x}_i &= \mathbf{y}_i\mathbf{z}_i \\
\mathbf{z}_i &\sim \text{Bernoulli}(\pi_i) \\
\mathbf{y}_i &\sim \text{N}(0, \Sigma) \\
\pi &\sim \text{Beta}(a, b)
\end{aligned}$$

where a and b are parameters for beta process, $\mathbf{x}$ is the signal drawn from the beta process and $\mathbf{z} \in \{0, 1\}$ is a vector of binary. $\mathbf{y}$ or Gaussian distribution decides the

magnitude of the element in $\mathbf{x}$; while $\mathbf{z}$ or the Bernoulli distribution controls the sparsity of $\mathbf{x}$. Thus beta process can be used to draw sparse vectors, which has been used in Zhou *et al.* (2009) for dictionary learning and shown several advantages over existing dictionary learning methods, .e.g., K-SVD Aharon *et al.* (2005).

Chapter 3

DISCRIMINATIVE K-SVD FOR DICTIONARY LEARNING IN FACE RECOGNITION

In a sparse-representation-based face recognition scheme, the desired dictionary should have good representational power (i.e., being able to span the subspace of all faces) while being able to support optimal discrimination of the classes (i.e., different human subjects). In this chapter, we describe the first problem and our contributions: how to learn an over-complete dictionary that attempts to simultaneously achieve the above two goals. The proposed method, discriminative K-SVD (D-KSVD), is based on extending the K-SVD algorithm by incorporating the classification error into the objective function, thus allowing the performance of a linear classifier and the representational power of the dictionary being considered at the same time by the same optimization procedure. The D-KSVD algorithm finds the dictionary and solves for the classifier using a procedure derived from the K-SVD algorithm, which has proven efficiency and performance. This is in contrast to most existing work that relies on iteratively solving sub-problems with the hope of achieving the global optimal through iterative approximation. We evaluate the proposed method using two commonly-used face databases, the Extended YaleB database and the AR database, with detailed comparison to 3 alternative approaches, including the leading state-of-the-art in the literature. The experiments show that the proposed method outperforms these competing methods in most of the cases. Further, using Fisher criterion and dictionary incoherence, we also show that the learned dictionary and the corresponding classifier is indeed better-posed to support sparse-representation-based recognition.

3.1 Introduction

Face recognition is a challenging computer vision task that has seen active research for many years Zhao *et al.* (2003). Well-known, conventional approaches include Eigenface Turk and Pentland (1991) and Fisherface Belhumeur *et al.* (1997), among others. These methods usually involve two stages: feature extraction and classification. Recently, a lot of attention has been given to applying sparse-representation-based techniques to computer vision and image processing problems, such as image denoising Elad and Aharon (2006), image inpainting Mairal *et al.* (2008c), image compression Bryt and Elad (2008a)Bryt and Elad (2008b). In particular, the SRC algorithm proposed in Wright *et al.* (2009b) uses sparse representation for face recognition: training face images are used to form a dictionary, and classifying a new face image is achieved through finding its sparse coefficients with respect to this dictionary. Unlike conventional methods such as Eigenface and Fisherface, SRC does not need an explicit feature extraction stage. The superior performance reported in Wright *et al.* (2009b) suggests that this is a promising direction for face recognition.

The basic way of forming the dictionary by using all the training images may result in a huge size for the dictionary, which is detrimental to the subsequent sparse solver. For example, we may have 32 images for each person (e.g., as in the Extended YaleB database Georghiades *et al.* (2001)). Then the number of atoms in the dictionary will be 32 times the number of people. Thus for a large face database with thousands of people, the sheer size of the dictionary becomes a practical concern. One may manually select a subset of the training images to be used in the dictionary, as done in Wright *et al.* (2009b). But this is not only tedious but also sub-optimal since there is no guarantee that the manually-selected images form the best dictionary. Methods for learning a small-sized dictionary for sparse-coding from the training data have been proposed recently. For example, the K-SVD algorithm Aharon *et al.* (2006) learns an over-complete dictionary from a set of signals. The algorithm has been shown to work well in image compression and denoising. K-SVD focuses on

only the representational power of the dictionary (or the efficiency of sparse coding under the dictionary) without considering its capability for discrimination. Another recent work Pham and Venkatesh (2008) attempts to address this issue by further iteratively updating the K-SVD-trained dictionary based on the outcome of a linear classifier, hence obtaining a dictionary that may be also good for classification in addition to having the representational power. Other efforts along similar direction include Mairal *et al.* (2008a) and Mairal *et al.* (2008b), which use more sophisticated objective functions in dictionary optimization in training stage in order to gain some discriminative power for the dictionary.

In this chapter, we propose to extend the K-SVD algorithm to learn an over-complete dictionary from a set of labeled training face images. By directly incorporating the labels in the dictionary-learning stage (as opposed to relying on iteratively updating the dictionary using feedback from the classification stage as in Pham and Venkatesh (2008)), we can efficiently obtain a dictionary that retains the representational power while making the dictionary discriminative (i.e., supporting sparse-coding-based classification). We also propose a corresponding classification algorithm based on the learned dictionary. Incorporating the classification stage directly into the dictionary-learning procedure has the potential of avoiding the local minimum that may be encountered more often in the approach of Pham and Venkatesh (2008), which computes the sub-optimal solution by alternating between solving subset of parameters while fixing others. Furthermore, the complexity of the proposed method is bounded by that of the K-SVD, while the approach of Pham and Venkatesh (2008) involves multiple additional optimization procedures.

To demonstrate the effectiveness and the advantage of the proposed method for face recognition, extensive experiments have been carried out using two commonly-used face databases: the extended YaleB database Lee *et al.* (2005) and the AR database Martinez and Benavente (2007). In addition to comparing the recognition rates of our method with those from existing state-of-the-art approaches, we also

analyze and compare the performance of the classifiers based on Fisher criterion. The learned dictionaries are also compared in terms of dictionary incoherence Candès *et al.* (2006)Donoho (2006). The experimental results show that the proposed method has some clear advantages. In particular, the experiments show that with the same dictionary size or with dictionaries of randomly-chosen training images, our method can obtain better recognition rate than the SRC algorithm.

The rest of this chapter is organized as follows. We first briefly describe in Section 3.2 the basic formulation of the problem of face recognition based on spare-representation using an over-complete dictionary. Then we present the related work in the literature in Section 3.3 followed by the proposed algorithm in Section 3.4. The experiments and analysis of the results are reported in Section 3.5. We concludes with discussion in Section 3.6.

3.2 Basic Formulation of the Problem

It has been observed that images of human face under varying illumination conditions and expressions lie on a special low-dimensional space Belhumeur *et al.* (1997) Basri and Jacobs (2003). In a sparse-representation-based face recognition scheme like the SRC algorithm, this observation is exploited for recognition through sparse-coding of a testing face image using an over-complete dictionary of the training faces. Our method follows this scheme, which we briefly outline in the below.

Given sufficient samples of the i-th person, $\mathbf{A}_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \ldots, \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m*n_i}$, any test sample $y \in \mathbb{R}^m$ from the same class will approximately lie in the subspace spanned by the training samples associated with same class:

$$\mathbf{y} = a_{i,1} * \mathbf{v}_{i,1} + a_{i,2} * \mathbf{v}_{i,2} + \cdots + a_{i,n_i} * \mathbf{v}_{i,n_i} \tag{3.1}$$

where $a_{i,j}$ is a scalar.

By grouping samples from all the classes, we form a dictionary $\mathbf{A}$:

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k] = [\mathbf{v}_{1,1}, \mathbf{v}_{1,2}, \ldots, \mathbf{v}_{k,n_k}] \tag{3.2}$$

where k is the number of classes. Then the linear representation of $\mathbf{y}$ can be written in terms of all samples as:

$$\mathbf{y} = a_{1,1} * \mathbf{v}_{1,1} + a_{1,2} * \mathbf{v}_{1,2} + \cdots + a_{k,n_k} * \mathbf{v}_{k,n_k} = \mathbf{A} * \mathbf{x}_0 \tag{3.3}$$

where $\mathbf{x}_0 = [0, , \ldots, 0, a_{i,1}, a_{i,2}, \ldots, a_{i,n_i}, 0, \ldots, 0] \in R^n$ is a vector of coefficient whose entries are all zero except for those associated with the i-th class.

Based on this idea, if we extract the coefficient $\alpha_0(j)$ associated with the j-th person and reconstruct the image as

$$\mathbf{y}(j) = \mathbf{D} * \alpha_0(j) \tag{3.4}$$

we can expect that the reconstruction error $e(j) = \|\mathbf{y} - \mathbf{y}(j)\|_2$ will be large for any general $j \neq i$ except for $e(i)$. We can use this idea to recognize the test sample. While such a scheme has been shown to be able to generate the state-of-art results in Wright *et al.* (2009b), there are a few practical drawbacks associated with the method. For example,

1. In order to improve the representational power of the dictionary, we need to use a large number of training samples for each person. But a large dictionary is detrimental for the subsequent sparse solver.

2. In order to ensure that the dictionary atoms can span the underlying subspace reasonably well, we need to carefully choose the training images. For example, in Wright *et al.* (2009b), for the AR database, the authors manually chose 7 normal images (without artificial disguise) from Section 1 for each person.

3.3 Related Work

The above drawbacks associated with the SRC algorithm may be overcome if we can learn a smaller-sized dictionary from the given training images while maintaining the representational power of the dictionary. For example, the K-SVD algorithm

Aharon *et al.* (2006) may be employed for this purpose, which finds a solution for the following problem:

$$< \mathbf{D}, \alpha >= \underset{\mathbf{D},\alpha}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{D} * \alpha\|_2 \ subject\ to\ \|\alpha\|_0 \leq T \tag{3.5}$$

where $\mathbf{Y}$ is the matrix of all input signals (the training face images in our case), and T is a parameter to impose the sparsity prior. In Eqn. 3.5, each column of $\mathbf{D}$ is normalized to have unit norm. This K-SVD formulation has been found to work well for real images in applications such as image denoising and face image compression. However, since the objective function in Eqn. 3.5 considers only the reconstruction error and the sparsity of the coefficient, the learned dictionary is not optimized for a classification task. In other words, the learned dictionary may not have the best discriminative power despite its representational power.

Efforts have been reported on improving a dictionary-learning procedure for classification tasks. In Mairal *et al.* (2008a) and Mairal *et al.* (2008b), an extra term was introduced to consider the classifier performance in dictionary learning. For a binary classifier, this term can be represented by

$$< \theta >= \underset{\theta}{\operatorname{argmin}} \sum_i C(\mathbf{h}_i * f(\alpha_i, \theta)) + \lambda_1 * \|\theta\|_2 \tag{3.6}$$

where θ is the parameter of the classifier, $\mathbf{h}_i$ is the label and $C(x)$ is logistic loss function $C(x) = log(1 + e^{-x})$. The resultant problem is very complex and thus there is no direct method to find the solution. Instead, projected gradient descent was used in finding approximate solutions in the paper.

Another example is Pham and Venkatesh (2008), which uses a simpler formulation for considering the classifier performance:

$$< \mathbf{W}, \mathbf{b} >= \underset{\mathbf{W},\mathbf{b}}{\operatorname{argmin}} \|\mathbf{H} - \mathbf{W} * \alpha - \mathbf{b}\ \|_2 + \beta' \|\mathbf{W}\|_2 \tag{3.7}$$

where $\mathbf{W}$, $\mathbf{b}$ are parameters for a linear classifier $\mathbf{H} = \mathbf{W} * \alpha + \mathbf{b}$. Each column of $\mathbf{H}$ is a vector: $\mathbf{h}_i = [0, 0, \ldots, 1 \ldots, 0, 0]$, where the position of non zero element indicates

the class. So $\|\mathbf{H} - \mathbf{W} * \alpha + \mathbf{b}\|_2$ is classification error and $\|\mathbf{W}\|_2$ is the regularization penalty term. We can set $\mathbf{b}$ to zero for simplicity.

Considering Eqn. 3.5 and Eqn. 3.7 at the same time, we can naturally define the following problem for learning a dictionary with both discriminative power and representative power:

$$\begin{aligned} < \mathbf{D}, \mathbf{W}, \alpha > \quad &= \quad \underset{\mathbf{D},\mathbf{W},\alpha}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{D} * \alpha\|_2 + \gamma * \|\mathbf{H} - \mathbf{W} * \alpha\|_2 + \beta * \|\mathbf{W}\ \|_2 \\ &s.t. \quad \|\alpha\|_0 \leq T \end{aligned} \tag{3.8}$$

where $\mathbf{Y}$ is the set of input signals, $\mathbf{D}$ the dictionary, α the coefficient, $\mathbf{H}$ the label of the training images, $\mathbf{W}$ the parameter of the classifier, and γ and β are scalars controlling the relative contribution of the terms to the overall objective function.

The above formulation may be viewed as a special case of Pham and Venkatesh (2008) without considering the unlabeled data therein. However, our emphasis is on viewing the formulation of Eqn. 3.8 as an extended K-SVD problem and thus the solution (to be presented in subsequent subsections) will be solved by a K-SVD-like algorithm. This is in contrast with the sophisticated (and computationally involving) optimization procedures used in Mairal *et al.* (2008a), Mairal *et al.* (2008b), and Pham and Venkatesh (2008). To better illustrate this point, we describe the following iterative procedure for solving the problem of Eqn. 3.8 (which we will refer to as the Baseline Algorithm later):

Baseline Algorithm

Initialize D and α with K-SVD method by Eqn. 3.5;

while not converged **do**

Calculate W in Eqn. 3.7 when D and α fixed;

Calculate α when D and W;

Calculate D when α and W fixed;

Check convergence;

end while

Essentially, the above procedure is effectively the algorithm of Pham and Venkatesh (2008), except that here we only consider labeled data. Hence this Baseline Algorithm will be used in our comparison of the proposed method with that of Pham and Venkatesh (2008).

3.4 Proposed Method

The Baseline Algorithm mentioned in Sec. 3.3 can only find an approximate solution to the problem of Eqn. 3.8, since the problem in Eqn. 3.8 is not convex and in each step of the method, it only finds solution for a sub-problem of Eqn. 3.8. While practically speaking, the final solution may converge to the real solution, the method has big potential of getting stuck at local minimum of the sub-problems. Additionally, as is obvious from the Baseline Algorithm, in each iteration there are three optimization problems involved and thus the convergence, if it happens, will be slow to reach. To get around these issues, and to leverage the proved performance of the K-SVD algorithm, we propose the following Discriminative K-SVD (D-KSVD) algorithm, which uses K-SVD to find the globally optimal solution for all the parameters simultaneously. The task is formulated as solving the following problem

$$\begin{aligned} < \mathbf{D}, \mathbf{W}, \alpha > \quad &= \quad \underset{\mathbf{D},\mathbf{W},\alpha}{\operatorname{argmin}} \| \begin{pmatrix} \mathbf{Y} \\ \sqrt{\gamma} * \mathbf{H} \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ \sqrt{\gamma} * \mathbf{W} \end{pmatrix} * \alpha \|_2 + \beta * \|\mathbf{W}\|_2 \\ & s.t. \quad \|\alpha\|_0 \leq T \end{aligned} \tag{3.9}$$

We adopt the protocol in the original K-SVD algorithm: the matrix $\begin{pmatrix} \mathbf{D} \\ \sqrt{\gamma} * \mathbf{W} \end{pmatrix}$ is always normalized column-wise. Therefore, we can further drop the regularization

penalty term $\|\mathbf{W}\|_2$, and thus the final formulation of the problem can be written as:

$$\begin{aligned} < \mathbf{D}, \mathbf{W}, \alpha > \quad &= \quad \underset{\mathbf{D},\mathbf{W},\alpha}{\operatorname{argmin}} \| \begin{pmatrix} \mathbf{Y} \\ \sqrt{\gamma} * \mathbf{H} \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ \sqrt{\gamma} * \mathbf{W} \end{pmatrix} * \alpha \|_2 \\ & s.t. \quad \|\alpha\|_0 \leq T \end{aligned} \tag{3.10}$$

Now, the problem of Eqn. 3.10 can be efficiently solved by updating the dictionary atom by atom with the following method: For each atom $\mathbf{d}_k$ and the corresponding coefficient α_k, we solve the following problem

$$< \mathbf{d}_k, \alpha_k >= \underset{\mathbf{d}_k,\alpha_k}{\operatorname{argmin}} \|\mathbf{E}_k - \mathbf{d}_k * \alpha_k\|_F \tag{3.11}$$

where $\mathbf{E}_k = \mathbf{Y} - \sum_{i \neq k} \mathbf{d}_i * \alpha_i$ and $\mathbf{Y}$ is the training data. $\|\cdot\|_F$ denotes the Frobenius norm. This is essentially the same problem that K-SVD has solved and thus the the solution to Eqn. 3.11 is given by

$$\begin{aligned} \mathbf{U} * \Sigma * \mathbf{V}^T &= \text{SVD}(\mathbf{E}_k) \\ \tilde{\mathbf{d}_k} &= \mathbf{U}(:, 1) \\ \tilde{\alpha_k} &= \Sigma(1, 1) * \mathbf{V}(1, :)^T \end{aligned} \tag{3.12}$$

where $\mathbf{U}(:, 1)$ denotes the first column of $\mathbf{U}$.

3.4.1 Algorithm for Classification

Upon the completion of training with the labeled data in the previous D-KSVD algorithm, we obtain an learned dictionary $\mathbf{D}$ and a classifier $\mathbf{W}$. However, the dictionary $\mathbf{D}$ does not readily support a sparse-coding based representation of a new test image, since $\mathbf{D}$ and $\mathbf{W}$ are normalized jointly in the previous learning algorithm, i.e,

$$\| \begin{pmatrix} \mathbf{d}_i \\ \sqrt{\gamma} * \mathbf{w}_i \end{pmatrix} \|_2 = 1 \tag{3.13}$$

Note that we cannot simply re-normalize $\mathbf{D}$ column-wise by itself, since in the training stage $\mathbf{W}$ is obtained with the original, un-normalized $\mathbf{D}$. Hence, we need to figure out

a way of obtaining a valid (normalized) dictionary and the corresponding classifier, based on the learning results, $\mathbf{D}$ and $\mathbf{W}$. To this end, we prove the following lemma which establishes the relationship between the desired ($\mathbf{D}'$, $\mathbf{W}'$) and the learned ($\mathbf{D}$, $\mathbf{W}$).

Lemma: The normalized dictionary $\mathbf{D}'$ and the corresponding classifier $\mathbf{W}'$ can be computed as

$$\begin{aligned} \mathbf{D}' &= \{\mathbf{d}'_1, \mathbf{d}'_2, \ldots, \mathbf{d}'_k\} = \{\frac{\mathbf{d}_1}{\|\mathbf{d}_1\|_2}, \frac{\mathbf{d}_2}{\|\mathbf{d}_2\|_2}, \ldots, \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_2}\} \\ \mathbf{W}' &= \{\mathbf{w}'_1, \mathbf{w}'_2, \ldots, \mathbf{w}'_k\} = \{\frac{\mathbf{w}_1}{\|\mathbf{d}_1\|_2}, \frac{\mathbf{w}_2}{\|\mathbf{d}_2\|_2}, \ldots, \frac{\mathbf{w}_k}{\|\mathbf{d}_k\|_2}\} \end{aligned} \tag{3.14}$$

where $\mathbf{d}_i$ and $\mathbf{w}_i$ denote the i-th column of $\mathbf{D}$ and $\mathbf{W}$, respectively.

Proof: If $\mathbf{y}$ is a vectorized image, then

$$\begin{aligned} \mathbf{y} &= \mathbf{D} * \alpha = \sum_i \alpha_i * \mathbf{d}_i = \sum_i \alpha_i * \|\mathbf{d}_i\|_2 * \frac{\mathbf{d}_i}{\|\mathbf{d}_i\|_2} = \sum_i \alpha'_i * \mathbf{d}'_i = \mathbf{D}' * \alpha' \\ \mathbf{l} &= \mathbf{W} * \alpha = \sum_i \alpha_i * \mathbf{w}_i = \sum_i \alpha_i * \|\mathbf{d}_i\|_2 * \frac{\mathbf{w}_i}{\|\mathbf{d}_i\|_2} = \sum_i \alpha'_i * \mathbf{w}'_i = \mathbf{W}' * \alpha' \end{aligned} \tag{3.15}$$

where $\mathbf{d}'_i = \frac{\mathbf{d}_i}{\|\mathbf{d}_i\|_2}$ and $\mathbf{w}'_i = \frac{\mathbf{w}_i}{\|\mathbf{d}_i\|_2}$ are the i-th column of $\mathbf{D}$ and $\mathbf{W}$ respectively.

With the normalized $\mathbf{D}'$, we can find the sparse coefficients for a given test image $\mathbf{y}$ by solving the following problem

$$< \alpha' >= \underset{\alpha'}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}' * \alpha'\|_2 + \sigma * \|\alpha'\|_0 \tag{3.16}$$

This is the typical sparse-coding problem and in practice we often resort to the following convex optimization problem

$$< \alpha' >= \underset{\alpha'}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}' * \alpha'\|_2 + \sigma * \|\alpha'\|_1 \tag{3.17}$$

which can be solved by many ℓ_1 optimization methods, such as GPSR Figueiredo *et al.* (2007), L1 magic Candes and Romberg (2005) and so on. The stability of the solution depends on the incoherence of $\mathbf{D}'$ and sparsity of α' Tropp (2006). When α' is sparse and $\mathbf{D}'$ is sufficiently incoherent, Orthonormal Matching Pursuit Chen *et al.* (1991) can also find the sparse coefficient Figueiredo *et al.* (2007). According to our

experiments with large face databases, OMP works well and run faster than other L1-optimization methods mentioned above. Thus the results reported in this paper are based on the OMP method.

The final classification of a test image is based on its sparse coefficient α', which carries most discriminative information. We can simply apply the linear classifier $\mathbf{W}'$ to α' and obtain the label of the image:

$$\mathbf{l} = \mathbf{W}' * \alpha' \tag{3.18}$$

where $\mathbf{l}$ is a vector.

Note that the coefficient α' can be viewed as the weight of each atom in reconstructing the test image. Thus we can view each column $\mathbf{w}'_k$ of $\mathbf{W}'$ as a factor for measuring the similarity of atom $\mathbf{d}'_k$ to each class. Therefore, $\mathbf{l} = \mathbf{W}' * \alpha'$ is the weighted similarity of the test image $\mathbf{y}$ to each class. In this sense, the label of test image $\mathbf{y}$ is decided by the index i where $\mathbf{l}_i$ is the largest among all elements of the $\mathbf{l}$ computed in Eqn. 3.18. Obviously, in the ideal case, $\mathbf{l}$ will be of the form $\mathbf{l} = \{0, 0, \ldots, 1, \ldots, 0, 0\}$ (i.e, with only one non-zero entry, which equals to 1).

3.5 Experiments and Analysis

In this section, we first use a simulation experiment (still based real face images) to compare the proposed D-KSVD method with the method of Pham and Venkatesh (2008). (As there is no code publicly available for the method in Pham and Venkatesh (2008), our comparison is based on our implementation of the Baseline Algorithm discussed in Sect. 3.1, which is essentially the same as that of Pham and Venkatesh (2008).) Then we evaluate our method on two commonly-used face databases: the Extended YaleB database and the AR database. For comparison purpose, we also implemented the SRC algorithm. To gain more insights into how the proposed D-KSVD method may gain over a plain K-SVD technique, we also implemented an algorithm that directly uses the dictionary learned by the original K-SVD algorithm for face recognition. The training stage of this algorithm runs as follows:

1. Train $\mathbf{D}$ with K-SVD according to Eqn. 3.5;

2. Train $\mathbf{W}$ with equation $\mathbf{W} = (\alpha^T\alpha + \beta' * \mathbf{I})^{-1} * \alpha * \mathbf{H}^T$.

In this algorithm, $\mathbf{D}$ and $\mathbf{W}$ are trained independently. The test stage is done similarly to what described in Sect. 3.4.1. For simplicity, we will refer to this method simply as K-SVD thereafter.

All the experiments were run on Matlab 2008a. The PC we used has an Intel P4 2.8GHz CPU and 1 GB RAM.

3.5.1 Simulation Experiments

We used 52 images from 2 random persons in the AR database (26 images each person) for this simulation. These images contain all the possible conditions in the AR database: varying expressions, varying illumination, and different occlusion conditions. We used the same parameters in running the two competing methods: the proposed and the Baseline Algorithm (or the method of Pham and Venkatesh (2008)).

First, we compare the methods based on the Fisher criterion, which is commonly used to evaluate the performance of classifiers. Fisher criterion measures the ratio of between-class variance and in-class variance. A bigger value usually means a better classification result. For a two-class problem, the Fisher criterion can be computed as follows:

$$S = \frac{\|\mu_1 - \mu_2\|_2^2}{\frac{1}{C_1} * \sum_{i=1}^{C_1} \|\mathbf{x}_1(i) - \mu_1\|_2^2 + \frac{1}{C_2} * \sum_{i=1}^{C_2} \|\mathbf{x}_2(i) - \mu_2\|_2^2} \tag{3.19}$$

where, μ is the mean of the data, the subscripts are the class labels, and $\mathbf{x}$ is the data, which is the $\mathbf{l}$ computed in Eqn. 3.5 in this analysis (since we wanted to see how well the $\mathbf{l}$'s computed by the two methods are). We have visualized $\mathbf{l}$ for all 52 images in Fig. 3.1. The computed Fisher criteria of the two methods are listed in Table 3.1. It shows that our method get a bigger value for fisher criterion, which means our method gets a better classification result than the method in Pham and Venkatesh (2008) does.

Table 3.1: The result for fisher criterion in simulation experiments.

Method	D-KSVD	Baseline
Fisher Criterion	1.2431	1.0924

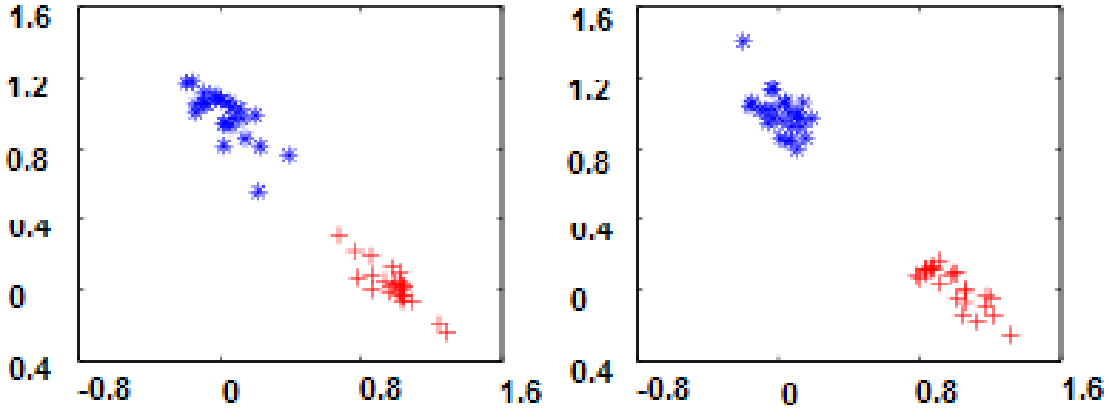


Figure 3.1: Visualizing the computed **l**. The plot on the left is from the Baseline Algorithm (or Pham and Venkatesh (2008)). The plot on the right is from the proposed D-KSVD method.

Second, we measure the incoherence of the dictionary which is critical for sparse representation. Y. Sharon *et al.* Sharon *et al.* (2007) proposed Equivalence Breakdown Point (EBP) for measuring the incoherence of the dictionary. However, computing EBP is computationally prohibitive for large dictionaries (e.g., of the size 600×500 as in our experiments). Thus we used the correlation coefficients from pairs of the atoms in the dictionary instead. This is calculated as

$$R(\mathbf{x}, \mathbf{y}) = \frac{cov(\mathbf{x}, \mathbf{y})}{\sqrt{cov(\mathbf{x}, \mathbf{x}) * cov(\mathbf{y}, \mathbf{y})}} \tag{3.20}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two atoms in the dictionary, *cov* computes the co-variance. A smaller coefficient R between two atoms means that they are more incoherent. Ideally, we want to have a small R for all possible pairs from the learned dictionary. We computed the largest R from all pairs in the two dictionaries learned from the Baseline Algorithm and the proposed D-KSVD algorithm, as reported in Table 3.2, which shows that the proposed method learns a better dictionary.

Table 3.2: The maximal pair-wised correlation for dictionary learned by D-KSVD and Baseline in simulation experiment.

Method	D-KSVD	Baseline
Max R value	0.7633	0.7830

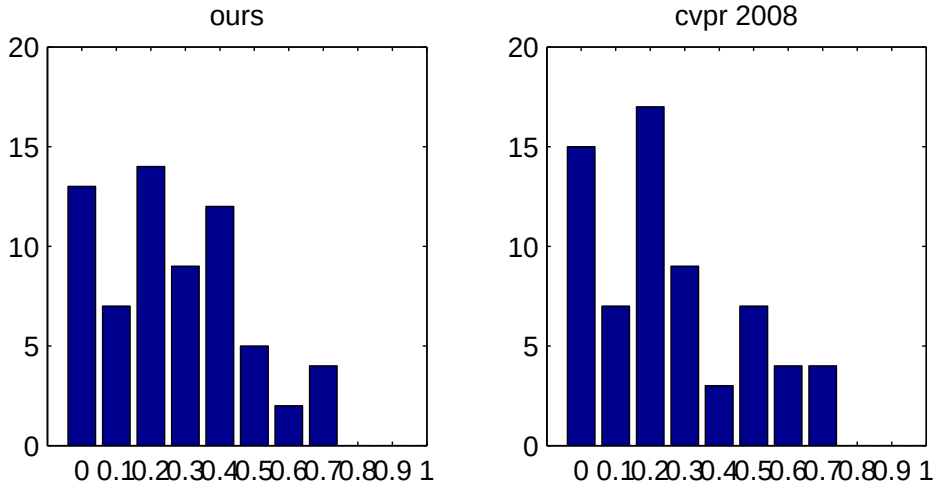


Figure 3.2: The histogram for pairwise correlation coefficient of the atoms in dictionary. The left one is dictionary with our method. The right one is that with Baseline

3.5.2 Results with the Extended YaleB Database

The Extended YaleB database contains about 2414 frontal face images of 38 individuals. Following Wright *et al.* (2009b), we used the cropped and normalized face images of 192*168 pixels, which were taken under varying illumination conditionsLee *et al.* (2005). We randomly split the database into two halves. One half, which contains 32 images for each person, was used for training the dictionary. The other half was used for testing. Further, we projected the face image $\in \mathbb{R}^{192*168}$ into a vector $\in \mathbb{R}^{504}$ with a randomly generated matrix, which is called Randomface Wright *et al.* (2009b). The learned dictionary contains 304 atoms, which corresponds to, on average, roughly 8 atoms for each person (but we must point out that, unlike in the SRC algorithm, in our method there is no explicit correspondence between the atoms and the labels of the people, since all the information is encapsulated into the discrimina-

tive dictionary and the corresponding classifier). The sparsity prior assumed in the learning was set to $T = 16$.

With this database, we tested 4 methods: SRC, K-SVD (as defined earlier in the beginning of Section 4.1), the Baseline Algorithm (or equivalently the method of Pham and Venkatesh (2008)), and the proposed D-KSVD method. The best result reported for SRC is 98.26% when there are 32 images per person in the dictionary. We also tested the performance of SRC when the dictionary is smaller (8 atoms per people). This set of results are denoted by $SRC\dagger$ in the subsequent tables. In Pham and Venkatesh (2008), the authors only used a few images (at most 4) per person for training and the recognition result was very poor (about 66.4%). For a fair comparison, we tested the Baseline Algorithm (essentially the method of Pham and Venkatesh (2008)) with more training images. In short, the key learning parameters used in the four methods were kept to be the same in our experiments.

All the results are summarized in Table 3.3. In the experiments, the scalar β and γ were set to 1. From the experiments, we found that most of the failure cases (about 46 out of 54) are from images under extreme illumination conditions. Some examples of these cases are given in Fig. 3.3. Thus, we performed another set of experiments with these "bad" images excluded (13 for each person). This was intended to show the true performance of the competing methods without the interference of images of extremely bad quality. The results of these new round experiments are listed in the last row of the table. From the table, it is clear that the proposed D-KSVD method always obtains better results than the K-SVD method and the Baseline (or the method of Pham and Venkatesh (2008)); In addition, for dictionaries of the same size, our method performs better than the SRC method.

We also evaluated the incoherence of the learned dictionaries in the experiments by calculating the correlation coefficient for each pair of atoms in the dictionary. Since the experiments involve multiple classes, and thus the correlation of the atoms may exhibit more complex patterns. To avoid the situation that a big R from a single

Table 3.3: The performance of the algorithms (recognition rate in %) for the Extended YaleB database. The 2nd row is the result when we used all 64 images for each people. The 3rd row is the result when we excluded 13 poor-quality images for each person. The images for training the dictionary were randomly selected.

D-KSVD	SRC	SRC†	K-SVD	Baseline
95.56	99.31	80.76	93.17	93.17
99.58	99.72	93.85	99.30	98.89

Table 3.4: The time for classifying one test image using the SRC method and the D-KSVD method on the extended YaleB database. We record the time for all the test images and then divide it by the number of images. The value is the average over 4 rounds. The unit is millisecond. The 2nd row is the result when we use all 64 images for each people. The 3rd row is the result when we use 51 images for each people.

Method	D-KSVD	SRC	SRC†
Case 1	84	120	83
Case 2	78	121	82

pair of atoms overshadows the correlation of all other pairs, in this case, we plot the histogram of the correlation coefficients. The results are given in Fig. 3.4 and Fig. 3.5 respectively. From these plots, it was found that the proposed D-KSVD method was able to generate a dictionary that contains more less-correlated atom pairs. That is, in the plots, the bars from the proposed method are on average slightly taller towards the left side of the axis of the correlation coefficients. (Probably one cannot expect to see dramatic difference in these plots, given that the performances of the algorithms are already very close and the improvement is at most a couple of percents. However, these few last percents are the hardest to obtain.)

In addition to the classification performance of the D-KSVD method and the SRC method, we also compared their speed performance for classifying one test image. We recorded the total time for classifying all the test images, and then divided it by the number of the test images, hence obtaining the average processing time for each testing image. We ran this for 4 rounds and calculated the average result, as shown in Table 3.4. From the results in Table 3.4, we can see that, with a smaller dictionary

Figure 3.3: Sample images under extreme illumination conditions. The left two are from one person and the right two from another person.

(304 atoms in the dictionary for D-KSVD and SRC†, and 1216 atoms in the dictionary for SRC), we can save about 1/3 of the time in testing. With a database involving more people, we can expect a smaller dictionary can save even more time (see the results below for the AR database too).

3.5.3 Results with the AR database

The AR database contains over 4000 color images for 126 people. For each person, there are 26 images taken in two different sections. These images contain 3 different illumination conditions, 3 different expressions and 2 different facial disguises (with sunglasses and scarf respectively). Thus this is a more challenging dataset. In our experiments, we used 2600 images from 50 male and 50 female. For each person, we randomly selected 20 images for training the dictionary and the other 6 for testing, which generally contain all the possible variations in the database. The results reported in the subsequent table are from the average of three such random spits of the training and testing images. The learned dictionary contains 500 atoms, i.e., roughly

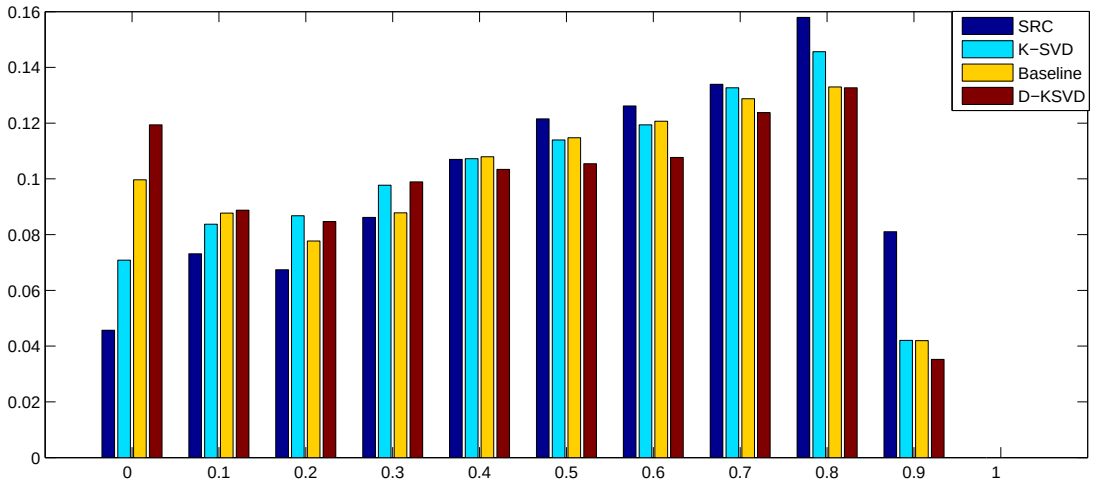


Figure 3.4: The histogram of pairwise correlation coefficients of the atoms in the learned dictionary. The dictionaries were trained by 4 different methods with the extended YaleB database.

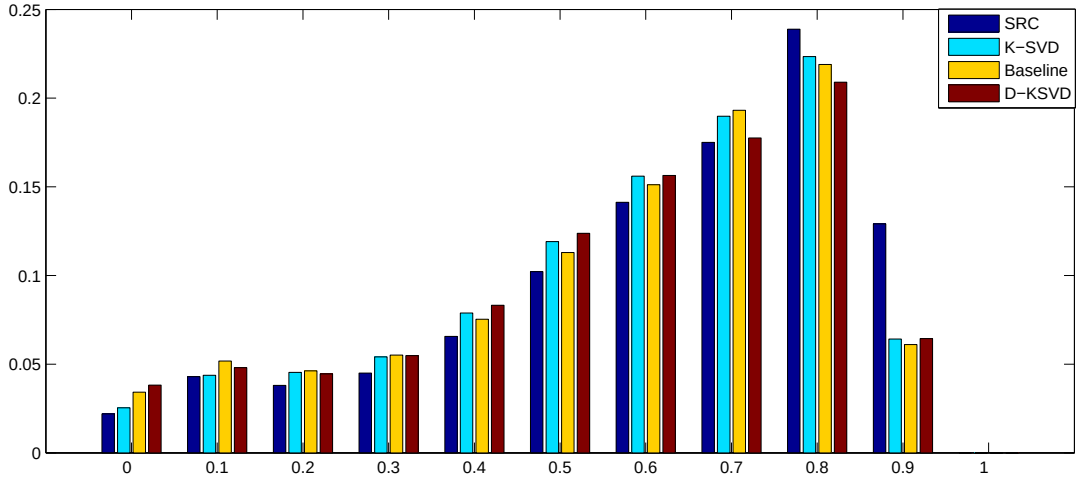


Figure 3.5: The histogram of pairwise correlation coefficients of the atoms in learned dictionary. The dictionaries were trained by 4 different methods with the extended YaleB database with 13 poor-quality images excluded.

5 atoms per person (but again, as discussed earlier, in our method there is no explicit correspondence between the atoms and the people). The sparsity prior was set to $T = 10$.

For direct comparison, we quoted the performance of the SRC algorithm on this dataset from Wright *et al.* (2009b), as listed in Table 3.5. It is worth noting that, in their experiments, they manually selected 7 images without facial disguise for each people from first section to build the dictionary. In our experiments, we tested

Table 3.5: The result reported for SRC. SRC‡ means first breaking the images into blocks, then classifying each block, finally using voting policy to decide image's label.

Test images	Without disguise	Sunglasses	Scarves
SRC	94.7%	87.0%	59.5%
SRC‡	NA	97.5%	93.5%

Table 3.6: The performance of the algorithms (recognition rate in %) for the AR database. SRC_n means there are n atoms per person in the dictionary learned by SRC. All other 3 methods use 500 atoms (roughly 5 per person). The images for training the dictionary were randomly selected.

D-KSVD	SRC_{20}	SRC_5	K-SVD	Baseline
95.0	90.50	68.14	88.17	93.0

the SRC algorithm with randomly-selected images for building the dictionary and experimented with two different dictionary sizes. We also tested K-SVD and the method of Pham and Venkatesh (2008) (the Baseline) on the AR database. In the experiments, all four methods use the same parameters. We also did the random projection as described earlier and in this case, this was from face images $\in \mathbb{R}^{165*120}$ into vectors $\in \mathbb{R}^{540}$.

The final results from all the methods are listed in Table 3.6. The experiments show that the performance of the SRC algorithm degraded dramatically when the training of the dictionary was based on randomly-selected images: when there are 5 images per person in the dictionary, the result is merely 68.14%. From the table, the proposed method outperforms all the competing methods.

As in the experiments with the Extended YaleB database, we also compared the speed performance of the D-KSVD method and the SRC method for classifying one test image on the AR database. The same method was used here. The result is shown in Table 3.7. For SRC_5 and D-KSVD, the dictionary has 500 atoms, and the size is 2000 atoms for SRC_{20}. As expected, for a database involving more people, a smaller dictionary can save more time, which is about 1/2 from the table.

Table 3.7: The time for classifying one test image using the SRC method and the D-KSVD method on AR database. We record the time for all the test images and then divide it by the number of images. The value is the average over 4 rounds. The unit is millisecond.

Method	D-KSVD	SRC_{20}	SRC_5
Result	62	131	76

3.6 Conclusion

We proposed a dictionary-learning approach, Discriminative K-SVD (D-KSVD), for face recognition. By adding a discriminative term into the objective function of the original K-SVD algorithm, we can ensure that the learned over-complete dictionary is both representative and discriminative. The solution of the new formulation follows a procedure derived from the original K-SVD algorithm and thus can be efficiently solved. Unlike existing approaches that iteratively solve sub-problems in order to approximate a global solution, our method directly finds all the parameters (the dictionary and the classifier) simultaneously. With extensive experiments on two large, commonly-used face databases, we demonstrated the advantages of the proposed method. The experimental results shows that: under the same learning condition, our method always outperforms K-SVD and the method of Pham and Venkatesh (2008); with the same dictionary size or with randomly chosen training images, our method outperforms the SRC algorithm. Our future work includes exploring both theoretically and empirically the structure of the learned dictionaries from our method and the competing methods, so as to reveal deeper insights on how to incorporate label information into dictionary learning. More extensive analysis on the speed performance of the algorithms is also another direction of interest and of practical importance.

Chapter 4

MINING DISCRIMINATIVE COMPONENTS WITH LOW-RANK AND SPARSITY CONSTRAINTS FOR FACE RECOGNITION

This chapter presents our second problem, considering the physical process of image formation, how can we extract a compact and effective subspace for face images. To crack this problem, we propose a novel image decomposition approach for an ensemble of correlated images, using low-rank and sparsity constraints. Each image is decomposed as a combination of three components: one common component, one condition component, which is assumed to be a low-rank matrix, and a sparse residual. For a set of face images of N subjects, the decomposition finds N common components, one for each subject, K low-rank components, each capturing a different global condition of the set (e.g., different illumination conditions), and a sparse residual for each input image. Through this decomposition, the proposed approach recovers a clean face image (the common component) for each subject and discovers the conditions (the condition components and the sparse residuals) of the images in the set. The set of $N + K$ images containing only the common and the low-rank components form a compact and discriminative representation for the original images. We design a classifier using only these $N + K$ images. Experiments on commonly-used face data sets demonstrate the effectiveness of the approach for face recognition through comparing with the leading state-of-the-art in the literature. The experiments further show good accuracy in classifying the condition of an input image, suggesting that the components from the proposed decomposition indeed capture physically meaningful features of the input.

4.1 Introduction

Face recognition has been an active research field for a few decades, and its challenges and importance continue to attract efforts from many researchers, resulting in many new approaches in recent years. The most recent literature may be divided into roughly two groups, where methods in the first group try to model the physical processes of image formation under different conditions (e.g., illumination, expression, pose etc.). For example, the approach of Lee *et al.* (2005) models the face image under varying illumination conditions to be a linear combination of images of the same subject captured at 9 specially designed illumination conditions; the SRC algorithm of Wright *et al.* (2009b) further assumes that face images with illumination and expression conditions can be represented as a sparse linear combination of the training instances (i.e., the dictionary atoms). On the other hand, the second group of approaches utilizes mathematical/statistical tools to capture the latent relations among face images for classification. E.g., the SUN approach Kanan and Cottrell (2010) uses the statistics of the human fixation of the images to recognize the face images, Volterrafaces Kumar *et al.* (2009b) finds a latent space for face recognition, where the ratio of intra-class distance over inter-class distance is minimized. One major advantage of the techniques in the first class comes from their being generative in nature, which allows these methods to accomplish tasks like face relighting or novel pose generation in addition to recognition. The second group of methods in a sense ignores the physical property of the faces images and treats them as ordinary 2D signals.

Although the methods in the first group have the above nice property, a baseline implementation usually requires dictionaries with training images as atoms and thus may face the scalability issue in real-world applications with a huge number of subjects. Hence efforts have also been devoted to reducing the size of the dictionary while attempting to retain the level of performance of the original dictionary. Examples include those that generate more compact dictionaries through some learning proce-

dure (e.g., Mairal *et al.* (2008a)) and those that attempt to extract subject-specific features that are effectively used as dictionary atoms (e.g., Nagesh and Li (2009)). Our approach belongs to the second group. Since the expressive power of the original dictionary-based techniques comes from largely the number of training images for each subject, a compact dictionary may suffer from degraded performance unless the reduced dictionary properly captures the conditions of the original data that are critical for a recognition task. For example, the method of Nagesh and Li (2009), while shown to be effective for expression-invariant recognition, is difficulty to generalize to handle global conditions such as illumination change, which often introduce to the data non-sparse conditions that cannot be captured by the sparsity model proposed therein.

Recognizing that non-sparse conditions such as illumination change and large occlusion are critical for face recognition, and that for a typical application we may assume only a finite number of such conditions (e.g., a relatively small number of illumination conditions or other conditions), in this chapter, we propose a model for representing a set of face images by decomposing them into three components: a common component shared by images of the same subject, a low-rank component capturing non-sparse global changes, and a sparse residual component. Such decomposition is partially inspired by the observation that the reconstruction of the image with the top few singular values and the corresponding singular vectors often capture the global information of the image, which can be represented by a low-rank matrix. To this end, a generic algorithm is proposed, with theoretic analysis on the convergence and parameter selection. The learned common and low-rank components form a compact and discriminative representation of the original set of images. A classifier is then built based on comparison of subspaces spanned by these components and by a novel image to be classified. This is very compact compared with the number of atoms in an over-determined dictionary such as that in Wright *et al.* (2009b). Further, by explicitly modeling non-sparse conditions, the proposed approach is able to

handle both illumination changes and large occlusions, which would fail methods like Nagesh and Li (2009).

To demonstrate the effectiveness of the proposed method, we first design synthetic experiments with known ground truth to verify its key capability in recovering the underlying common, low-rank and sparse components. Then we report results on three commonly-used data sets of real face images: the Extended YaleB dataset Georghiades *et al.* (2001), the CMU PIE dataset Sim *et al.* (2002) and the AR dataset Martinez and Benavente (1998).The experiments show that, the proposed approach obtained better performance than the SRC algorithm Wright *et al.* (2009b), which utilizes a much larger dictionary, and the SUN approach Kanan and Cottrell (2010). The proposed approach also achieves comparable result to Volterrafaces, which is the current state-of-the-art in the literature for a few commonly-used data sets. In addition, the proposed approach can explicitly model the most important feature of the subject and the conditions in the dataset. Experiments also show that the proposed method is robust to situations where a non-trivial percentage of the training images is unavailable. Further, the capability of the proposed approach for classifying the type of condition that an input image is subject to is also demonstrated by extensive experiments. This suggests that the proposed decomposition is able to obtain physically meaningful and thus potentially discriminative components.

We introduce the proposed method in Section 4.2, including the proposed model, the learning algorithm and the classification method. The experiments are reported and analyzed in Section 4.3. We conclude in Section 4.4 with a summary of the work and brief discussion on future work.

In the presentation, we use $\{\mathbf{X}_{i,j}\}_{i=1,j=1}^{N,M}$ denotes a set of $N \times M$ matrices, with $\mathbf{X}_{i,j}$ as its $(i,j)_{th}$ member. We assume that N is the number of the subjects, and M

the number of images per subject[1]. Thus $\mathbf{X}_{i,j}$ refers to j_{th} image of the i_{th} subject. When there is no confusion, we also use $\mathbb{X}$ to denote the set $\{\mathbf{X}_{i,j}\}_{i,j=1}^{N,M}$.

4.2 Proposed Method

In this section, we first present the general formulation of the proposed model in Section 4.2.1, and then present our algorithm for obtaining the desired decomposition in Section 4.2.2 and analysis of its convergence in Sec. 4.2.3. With these, a face recognition algorithm is then designed in Section 4.2.4.

4.2.1 Decomposing a Face Image Set

In many applications of image and signal processing, we often consider a set of correlated signals as an ensemble. For efficient representation, a signal in the ensemble can often be viewed as a combination of a common component, which is shared among all the signals in the ensemble, and an innovation component, which is unique to this signal. Many benefits can be drawn from this decomposition of the ensemble, such as obtaining better compression rate and being able to extract more relevant features Bengio *et al.* (2009). In face recognition, all the face images, especially the subset corresponding to a subject, may be naturally viewed as forming such an ensemble of correlated signals. In a sense, a sparse-coding approach like SRC *implicitly* figures out the correlation of the images in the ensemble via the sparse coefficients under the dictionary of the training images.

In this work, we aim at developing a new representation of this ensemble so that the face recognition task can be better supported. In particular, considering the common challenges such as illumination conditions and large occlusions, we want to have a representation that can *explicitly* model such conditions. To this end, we

[1]For simplicity, we assume that each subject has the same number of images, which can always be achieved by using some blank images, a situation the proposed method can handle.

propose the following decomposition of face images $\mathbf{X}_{i,j}$ in the ensemble $\mathbb{X}$ as:

$$\mathbf{X}_{i,j} = \mathbf{C}_i + \mathbf{A}_j + \mathbf{E}_{i,j},\ \forall \mathbf{X}_{i,j} \in \mathbb{X} \tag{4.1}$$

where $\mathbf{C}_i$ is the common part for Subject i, $\mathbf{A}_j$ is a low-rank matrix, and $\mathbf{E}_{i,j}$ is a sparse residual.

One essential difference between the proposed method and Robust PCA (RPCA Wright *et al.* (2009a)), is that RPCA assumes the signals are linearly dependent, with some sparsely corrupted entries in the signals. As a result, they build a big matrix with each signal as a vector. The big matrix would naturally be low-rank (because of the assumed inter-image correlation), in addition to having a sparse set of entries. On the other hand, the proposed decomposition is partially inspired by the observation that the reconstruction of the image with first few singular values and the corresponding singular vectors often capture the global information of the image Liu *et al.* (2008), e.g., illumination conditions, structured patterns, which can be represented by a low-rank matrix. Here the low-rank constraint arises from certain physical conditions (rather than due to inter-image correlation), and it is imposed on each individual image. Accordingly, we represent images by matrices rather than vectors, unlike other methods like Wright *et al.* (2009b),Wright *et al.* (2009a). With this, we can expect that:

$\mathbf{C}_i$ is a matrix representing the common information of images for Subject i, i.e., the common components;

$\mathbf{A}_j$ is a low-rank matrix capturing the global information of the image $\mathbf{X}_{i,j}$, e.g., illumination conditions (Fig. 4.3), structured patterns (Fig. 4.1); and

$\mathbf{E}_{i,j}$ is a sparse matrix pertaining to image-specific details such as expression conditions or noise with sparse support in the images.

In this modeling, we have assumed M different low-rank matrices, which are responsible for M different global conditions such as illumination conditions or large

occlusions, and they are shared among the images of different subjects. However, images of each subject do not necessarily contain all the M conditions, as we will show in Sec. 4.2.2.

The above model can also be explained via the Retinex theory, in which image $\mathbf{I}$ can be represented as:

$$\mathbf{I}(p,q) = \mathbf{R}(p,q) \cdot \mathbf{L}(p,q) \tag{4.2}$$

where $\mathbf{R}(x,y)$ is the reflectance at location (x,y), which depends on the surface property, $\mathbf{L}(x,y)$ is the illumination, and $\cdot$ is element-wise product. Converting this into the logarithm domain, we have

$$log(\mathbf{I}) = log(\mathbf{R}) + log(\mathbf{L}) \tag{4.3}$$

The above equation indicates that we can represent the intensity of the face image as follows:

$$log(\mathbf{X}_{i,j}) = \mathbf{C}_i + \mathbf{A}_j + \mathbf{E}_{i,j},\ \forall \mathbf{X}_{i,j} \in \mathbb{X} \tag{4.4}$$

where $\mathbf{C}_i = log(\mathbf{R})$ captures the common property of the images for Subject i, $\mathbf{A}_j = log(\mathbf{L})$ captures the lighting conditions, and $\mathbf{E}_{i,j}$ captures the residual. This is a variant of the model in Eqn. 4.1, and is especially suitable for illumination-dominated datasets such as the extended YaleB dataset and the CMU-PIE dataset.

With the above decomposition, the entire dataset containing $N \times M$ images can be compactly represented by N common components and K low-rank components. If we extract the common component $\mathbf{C}_i$ for face images of Subject i under different conditions, we expect that this common component $\mathbf{C}_i$ represents the most significant feature of that subject. The set of all the learned low-rank components $\mathbb{A} = \{\mathbf{A}_j\}_{j=1}^{M}$ represents all possible global conditions of the images in the set. Hence we may use $\mathbb{A}$ and $\mathbf{C}_i$ to span the subspace of the face images for Subject i, where, in the ideal case, any face images of this subject should lie in, barring a sparse residual. This suggests that we can utilize the subspaces for face recognition by identifying which subspace a test image is more likely to lie in, which is detailed in Sec. 4.2.4.

4.2.2 An Algorithm for the Decomposition

Based on Eqn. 4.1, we formulate the decomposition task as the following constrained optimization problem, with an objective function derived from the requirement of decomposing a set of images into some common components, some low-rank matrices and the sparse residuals:

$$\begin{aligned} \mathbb{C}, \mathbb{A}, \mathbb{E} &= \underset{\mathbb{C},\mathbb{A},\mathbb{E}}{\operatorname{argmin}} \sum_{i,j} \|\mathbf{A}_j\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}\|_1 \\ & s.t. \quad \mathbf{X}_{i,j} = \mathbf{C}_i + \mathbf{A}_j + \mathbf{E}_{i,j},\ \forall \mathbf{X}_{i,j} \in \mathbb{X} \end{aligned} \tag{4.5}$$

where $\|\mathbf{A}_j\|_* = \sum_i \sigma_i(\mathbf{A}_j)$ is the nuclear norm, $\|\mathbf{E}_{i,j}\|_1 = \sum_{p,q} |\mathbf{E}_{i,j}(p,q)|$ is the ℓ_1 norm and $\mathbb{E} = \{\mathbf{E}_{i,j}\}_{i,j=1}^{N,M}$. Note that, unlike Wright *et al.* (2009a) where a set of images are stacked as vectors of a low-rank matrix, we do not convert the image to a vector in the decomposition stage.

To absorb the constraints into the objective function, we can reformulate Eqn. 4.5 with augmented Lagrange multiplier as:

$$\begin{aligned} \mathbb{C}, \mathbb{A}, \mathbb{E} &= \underset{\mathbb{C},\mathbb{A},\mathbb{E}}{\operatorname{argmin}} \sum_{i,j} \|\mathbf{A}_j\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}\|_1 \\ &+ \frac{\mu_{i,j}}{2}\|\mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{A}_j - \mathbf{E}_{i,j}\|_F^2 + < \mathbf{Y}_{i,j}, \mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{A}_j - \mathbf{E}_{i,j} > \end{aligned} \tag{4.6}$$

where $\mathbf{Y}_{i,j}$ is the Lagrange multiplier, $\lambda_{i,j}$ and $\mu_{i,j}$ are scalars controlling the weight of sparsity and reconstruction error accordingly. When μ is sufficiently large, Eqn. 6.11 is equivalent to Eqn. 4.5. It is worth pointing out that, while for clarity we have written only the expression for Subject i, the optimization is actually done for the entire set of images, since the low-rank components are deemed as been shared by all images.

To solve the problem of Eqn. 6.11, a block coordinate descent algorithm may be designed, with each iterative step solving a convex optimization problem Candes and Plan (2009)Wright *et al.* (2009a) for one of the unknowns. To this end, we first describe the following three sub-solutions that are needed in each iteration of such

an algorithm, which correspond to solving only one of the unknowns (blocks) while fixing others.

Sub-solution 1: For finding an optimal $E_{i,j}$ in the t-th iteration, where the problem can be written as

$$\mathbf{E}_{i,j} = \underset{\mathbf{E}_{i,j}}{\operatorname{argmin}} \lambda_{i,j} \|\mathbf{E}_{i,j}\|_1 + \frac{\mu_{i,j}}{2} \|\mathbf{X}^E_{i,j} - \mathbf{E}_{i,j}\|^2_F + < \mathbf{Y}_{i,j}, \mathbf{X}^E_{i,j} - \mathbf{E}_{i,j} > \tag{4.7}$$

with $\mathbf{X}^E_{i,j} = \mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{A}_j$. So we do the following update Hale *et al.* (2008):

$$\mathbf{E}_{i,j} = S_{\frac{\lambda}{\mu_{i,j}}}(\mathbf{X}^E_{i,j} + \frac{1}{\mu_{i,j}} \mathbf{Y}_{i,j}) \tag{4.8}$$

where $S_\tau(\mathbf{X}) = sign(\mathbf{X}) \cdot max(0, |\mathbf{X}| - \tau)$.

Sub-solution 2: For finding an optimal $\mathbf{A}_k$ in the t-th iteration, where the problem can be written as

$$\mathbf{A}_j = \underset{\mathbf{A}_j}{\operatorname{argmin}} \sum_i \|\mathbf{A}_j\|_* + \frac{\mu_{i,j}}{2} \|\mathbf{X}^A_{i,j} - \mathbf{A}_j\|^2_F + < \mathbf{Y}_{i,j}, \mathbf{X}^A_{i,j} - \mathbf{A}_j > \tag{4.9}$$

We use the singular value thresholding algorithm Cai *et al.* (2008); Goldfarb and Ma (2011):

$$\begin{aligned} \mathbf{U}\Sigma\mathbf{V}^T &\leftarrow \frac{\sum_i \mu_{i,j} \mathbf{X}^A_{i,j} + \mathbf{Y}_{i,j}}{\sum_i \mu_{i,j}} \\ \mathbf{A}_j &= \mathbf{U} S_\tau(\Sigma) \mathbf{V}^T \end{aligned}$$

with $\mathbf{X}^A_{i,j} = \mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{E}_{i,j}$ and $\tau = \frac{N}{\sum_i \mu_{i,j}}$.

Sub-solution 3: The solution to the problem of finding optimal $\mathbf{C}_i$

$$\underset{\mathbf{C}_i}{\operatorname{argmin}} \frac{\mu_{i,j}}{2} \sum_j \|\mathbf{X}^C_{i,j} - \mathbf{C}_i\|^2_F + < \mathbf{Y}_{i,j}, \mathbf{X}^C_{i,j} - \mathbf{C}_i > \tag{4.10}$$

where $\mathbf{X}^C_{i,j} = \mathbf{X}_{i,j} - \mathbf{A}_j - \mathbf{E}_{i,j}$, can be obtained directly (by taking derivatives of the objective function and setting to zero) as

$$\mathbf{C}_i = \frac{\sum_j \mathbf{Y}_{i,j} + \mu_{i,j} \mathbf{X}^C_{i,j}}{\sum_j \mu_{i,j}} \tag{4.11}$$

As alluded earlier, the images of any given subject may not range over all possible M conditions. This may be equivalently viewed as a problem where some images are missing for the subject. We now show how this can be addressed in a principled way. Assume that Ω is the set of (i,j) where $\mathbf{X}_{i,j}$ is available and $\bar{\Omega}$ is the complement of Ω. To deal with those missing entries, we only need to set $\mathbf{Y}_{i,j}$, $\mu_{i,j}$ and $\mathbf{X}_{i,j}$ to 0 for $(i,j) \in \bar{\Omega}$ in the initialization stage. In each iteration, we do not update $\mathbf{E}_{i,j}$ for $(i,j) \in \bar{\Omega}$. The proposed decomposition algorithm will automatically infer the missing images.

With the above preparation, we now propose the following Algorithm 1 to solve Eqn. 6.11:

Algorithm 1: Learning the Decomposition

Input: $\mathbb{X}$, Ω, N, M, ρ, λ and τ;

Output: $\{\mathbf{C}_i\}_{i=1}^{N}$, $\{\mathbf{A}_j\}_{j=1}^{K}$ and $\{\mathbf{E}_{i,j}\}_{i,j=1}^{N,M}$;

{Initialization} $t = 0$, $\mathbf{C}_i^0 = \mathbf{A}_j^0 = \mathbf{E}_{i,j}^0 = 0$; $\mathbf{Y}_{i,j}^0 = \frac{\mathbf{X}_{i,j}}{\|\mathbf{X}_{i,j}\|_F}$, $\mu_{i,j}^0 = \frac{\tau}{\|\mathbf{X}_{i,j}\|_F}$ for $(i,j) \in \Omega$;

$\mathbf{Y}_{i,j}^0 = 0$, $\mu_{i,j}^0 = 0$ for $(i,j) \notin \Omega$;

while not converged **do**

Solve $\mathbf{E}_{i,j}$ for $(i,j) \in \Omega$ by Sub-solution 1:

Solve $\mathbf{A}_j$ for $j = 1, 2, ..., M$ with Sub-solution 2;

Solve $\mathbf{C}_i$ for $i = 1, 2, ..., N$ using Sub-solution 3;

$\mathbf{Y}_{i,j}^{t+1} = \mathbf{Y}_{i,j}^t + \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^{t+1} - \mathbf{A}_j^{t+1} - \mathbf{E}_{i,j}^{t+1})$ for $(i,j) \in \Omega$;

$\mu_{i,j}^{t+1} = \mu_{i,j}^t \rho$ for $(i,j) \in \Omega$;

$t = t + 1$;

check convergence;

end while

where for convergence, we check $\frac{\sum_{i,j}\|\mathbf{X}_{i,j}-\mathbf{C}_i-\mathbf{A}_j-\mathbf{E}_{i,j}\|_F^2}{\sum i,j\|\mathbf{X}_{i,j}\|_F^2}$ and if it is small enough (e.g., 10^{-6}), we terminate the algorithm. λ, τ and ρ are three parameters specified in input, which are discussed in Sec. 5.3.1.

4.2.3 Convergence of the Algorithm

The convergence property of an iterative optimization procedure like the algorithm proposed above is critical to its usefulness. The Algorithm 1 has similar convergence property as the methods described in Lin *et al.* (2010), which are also augmented Lagrange multiplier based approaches. We can draw the following theorem:

Theorem 1 If $\sum_{t=1}^{\infty}\mu_{i,j}^{t+1}(\mu_{i,j}^{t})^{-2} < \infty$ and $\lim_{t\to\infty}\mu_{i,j}^{t}(\mathbf{E}_{i,j}^{t+1}-\mathbf{E}_{i,j}^{t}) = 0\ \forall i,j$, then Algorithm 1 will converge to the optimal solution for the problem of Eqn. 4.5.

The proof of Theorem 1 is included in the Appendix A.

4.2.4 Face Recognition Using the Decomposition

With the components in Eqn. 4.1 estimated from the previous algorithm, we now discuss how to classify a test image. Recognizing that the sparse residual captures only image-specific details that have not been absorbed by the common or the global condition, we discard the sparse residuals from the decomposition (training) stage and keep only the common and the low-rank components.

Ideally a face image from Subject i should lie in a subspace spanned by its common component $\mathbf{C}_i$ and the low-rank components $\mathbb{A}$. Therefore, we propose the following classification scheme based on comparing the distance between subspaces spanned by the training components and those spanned by replacing the training common by the test image $\mathbf{y}$. We first build the subspace $\mathbf{S}_i$ for subject i, which contains all the linear combinations of the images of Subject i under all conditions, i.e.,

$$\mathbf{S}_i = \{\mathbf{x}|\mathbf{x} = \sum_k w_k \times (\mathbf{c}_i + \mathbf{a}_j)\,\forall \mathbf{w} \in \mathbb{R}^M\} \tag{4.12}$$

where $\mathbf{c}_i$ and $\mathbf{a}_j$ is the vectorized form of $\mathbf{C}_i$ and $\mathbf{A}_j$ respectively. Subspace $\mathbf{S}_i$ can be sufficiently represented by a set of "basis", i.e., $\{\mathbf{c}_i + \mathbf{a}_j\}_{j=1}^{M}$. Accordingly, we can build the subspace $\mathbf{S}_y$ for the test image y as the follows:

$$\mathbf{S}_y = \{\mathbf{x} | \mathbf{x} = \sum_k w_k \times (\mathbf{y} + \mathbf{a}_j) \, \forall \mathbf{w} \in \mathbb{R}^M\} \tag{4.13}$$

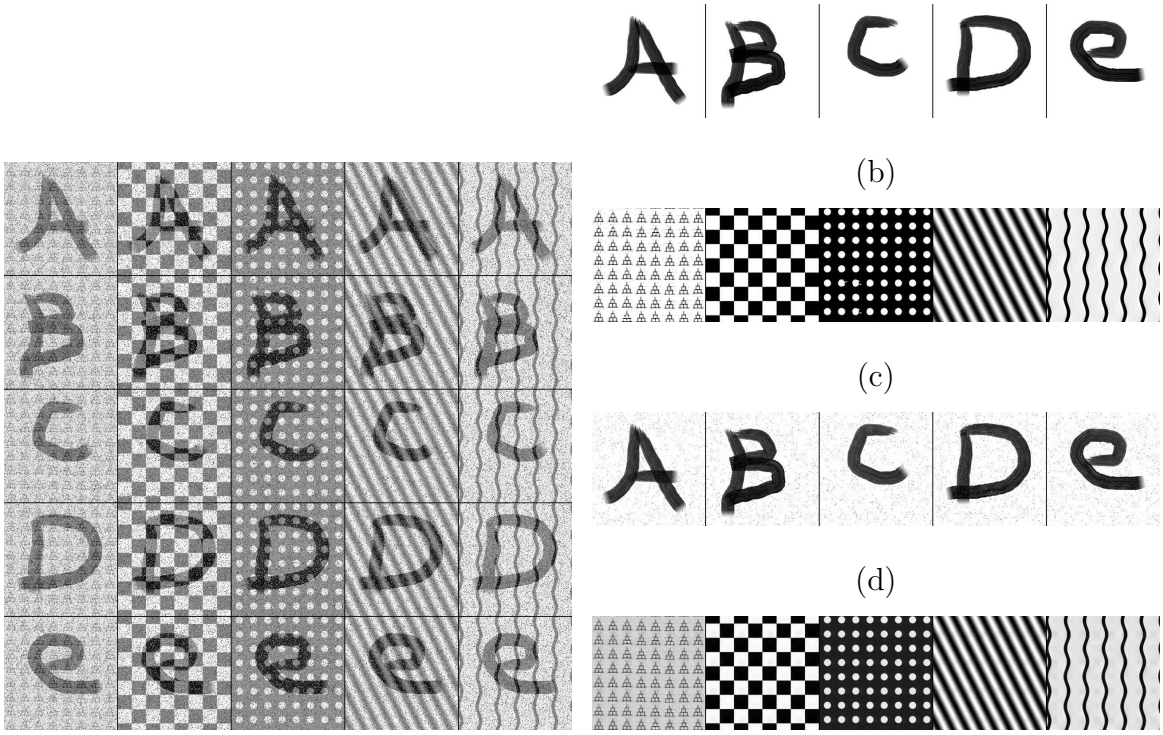
Then we use the principal angles Knyazev and Argentati (2002) between these subspace to measure their similarities. In this paper, the principal angles measure the cosine distance between the subspaces, which is calculated as $s(\mathbf{S}_i, \mathbf{S}_y) = \sum_k cos^2(\theta_k)$, where θ_k is the k_{th} principal angle between $\mathbf{S}_i$ and $\mathbf{S}_y$. The assign i as the label of $\mathbf{f}$, for which $s(\mathbf{S}_i, \mathbf{S}_y)$ is maximal.

4.3 Experimental Results

Experiments have been done to evaluate the proposed model and algorithms. In this section, we report several sets of results from such experiments. First, simulations (Sec. 5.3.1) are employed to demonstrate the convergence and parameter selection of the proposed decomposition algorithm. Then, we show the decomposition of the images from extended YaleB dataset and also how the learned components can be used to reconstruct new images in Sec. 4.3.2. Finally, we demonstrate the application of the proposed method and algorithms in classification tasks, including face recognition (Sec. 4.3.3) and identifying the conditions of the images (Sec. 4.3.4). The performance of the proposed method in face recognition task is compared with that of SRC Wright *et al.* (2009b), Volterrafaces Kumar *et al.* (2009b) and SUN Kanan and Cottrell (2010) on 2 commonly used datasets, i.e., extended YaleB Lee *et al.* (2005) and CMU-PIE Sim *et al.* (2002).

4.3.1 Simulation-based Experiments

In this subsection, we use synthetic data to demonstrate the convergence of the algorithm and selection of the parameters. The common components and condition components used in this experiment are shown in Fig. 4.1 (b,c), where the condition

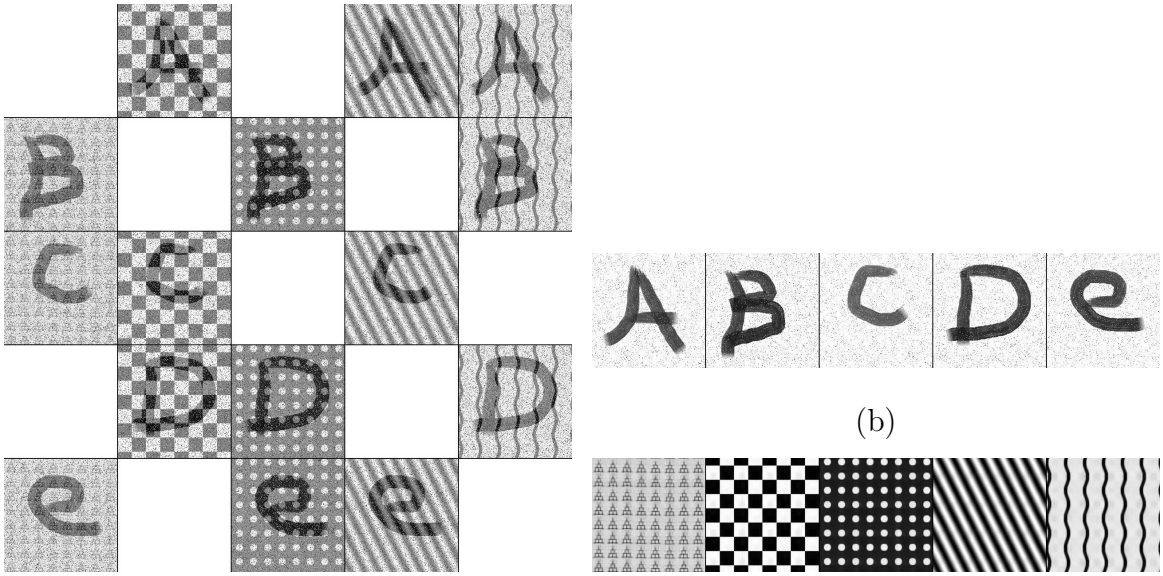


(a) (b) (c) (d) (e)

Figure 4.1: (a) shows the 25 images generated in the experiment, where the sparse part has 20% support of each image. (b,c) shows the ground truth of the common components and condition components accordingly. We also show the common components (d) and condition components (e) decomposed from (a) when $\rho = 1.25$ and $\tau = 2$.

components are from Portilla and Simoncelli (2000) and both components are rescaled to range $[0, 1]$. The sparse components are sampled from a uniform distribution in the range of $[0, 1]$. We use those components to generate 25 images, which are used in this experiment, as Eqn. 4.1.

Algorithm 1 in Sec. 4.2.2 requires three parameters, ρ controls the convergence speed; λ controls the sparsity of the sparse residuals; and τ is a scalar. In Lin *et al.* (2010), they suggest $\rho = 1.5$, $\tau = 1.25$ and $\lambda = \frac{1}{\sqrt{m}}$ for Robust PCA, where m is the width of $\mathbf{X}_{i,j}$. We have also found that $\lambda = \frac{1}{\sqrt{m}}$ is optimal from the experiments, thus we adopt this selection in our paper. From the experiment, we found that $\tau \in [1.25, 2]$ and $\rho = 1.25$ would be an optimal choice. Fig. 4.1 shows an example of the recovered



(a) (b) (c)

Figure 4.2: (a) the input data with 10 image manually removed, (b,c) is the common components and condition components decomposed from (a) accordingly.

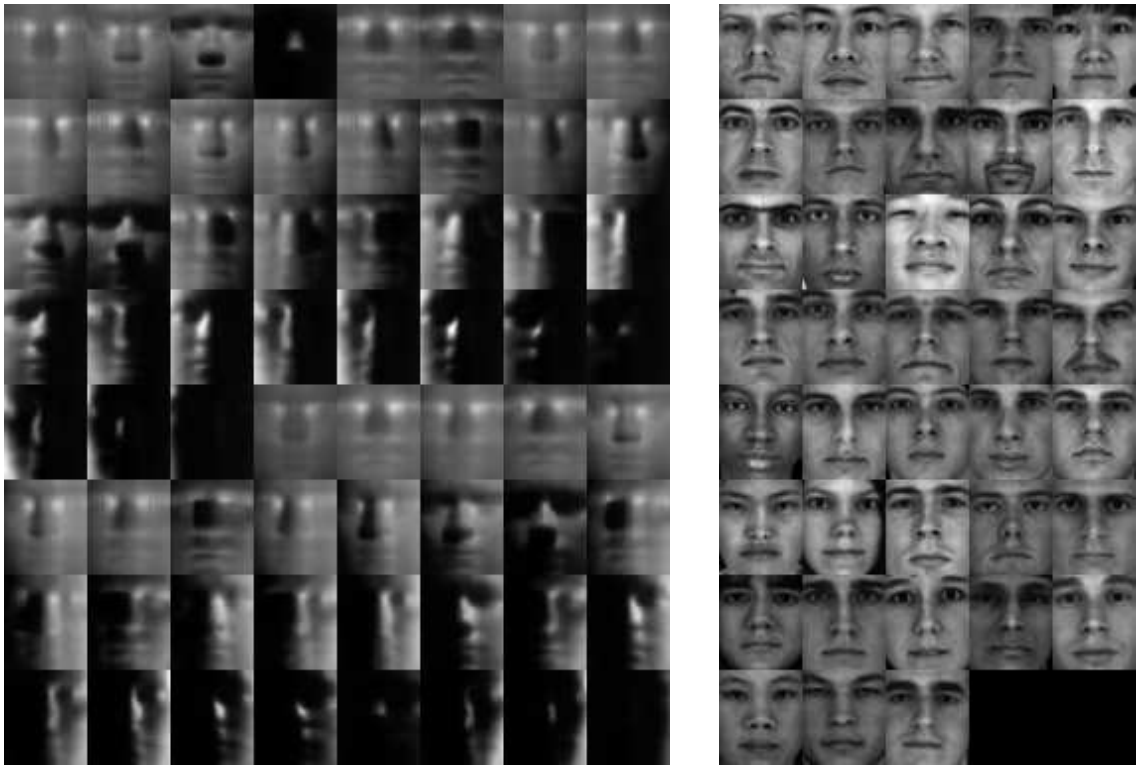
common components (d) and condition components (e) when the sparse part has 20% support of the image.[2]

To demonstrate the robustness of the algorithm, when only part of data is available, we randomly remove 10 images from the 25 images (Fig. 4.2(a)) and run the algorithm with the same set of parameters. The results are shown in Fig. 4.2, where (b) is the recovered common components and (c) is the recovered condition components. These results suggest that the algorithm is still able to produce reasonable results even with 40% of the images missing.

4.3.2 Decomposing a Set of Images

In this subsection, we first demonstrate the decomposition of the set of images from Extended YaleB datasetGeorghiades *et al.* (2001). All the 2432 images from 38 subjects under 64 illumination conditions were used. The common components and the condition components are illustrated in Fig. 4.3. Comparing these with the

[2]The recovered parts are subject to a linear shift and scaling. We identify the parameters for this linear shift and scaling then map them back with those parameters.



(a) (b)

Figure 4.3: The decomposition of the extended YaleB dataset. We use all the 2432 images which contain 38 subjects (b) and 64 illumination conditions (a).

original data, it is evident that the recovered commons are largely clean pictures of the subjects, while the condition components align well with the given illumination conditions. This experiment shows the capability of the proposed method with the Retinex model to discover the illumination conditions and the subject commons from a set of real images.

Next, we randomly pick 32 illumination conditions out of the decomposed 64 conditions and the common components of Subject 1 to form a subspace as described in Eqn. 4.14. Then we use the proposed method to identify whether a new image is in this subspace, by reconstructing this image as the linear combination of the "basis" of this subspace, i.e., $\mathbf{c}_1 + \mathbf{a}_j$. Fig. 4.4 shows an example, where the new image is also picked from Subject 1; and Fig. 4.5 shows another example, where the new image is picked from Subject 2. These examples suggest that the learned components can be used for identifying which subject a new image belongs to. Similarly, the

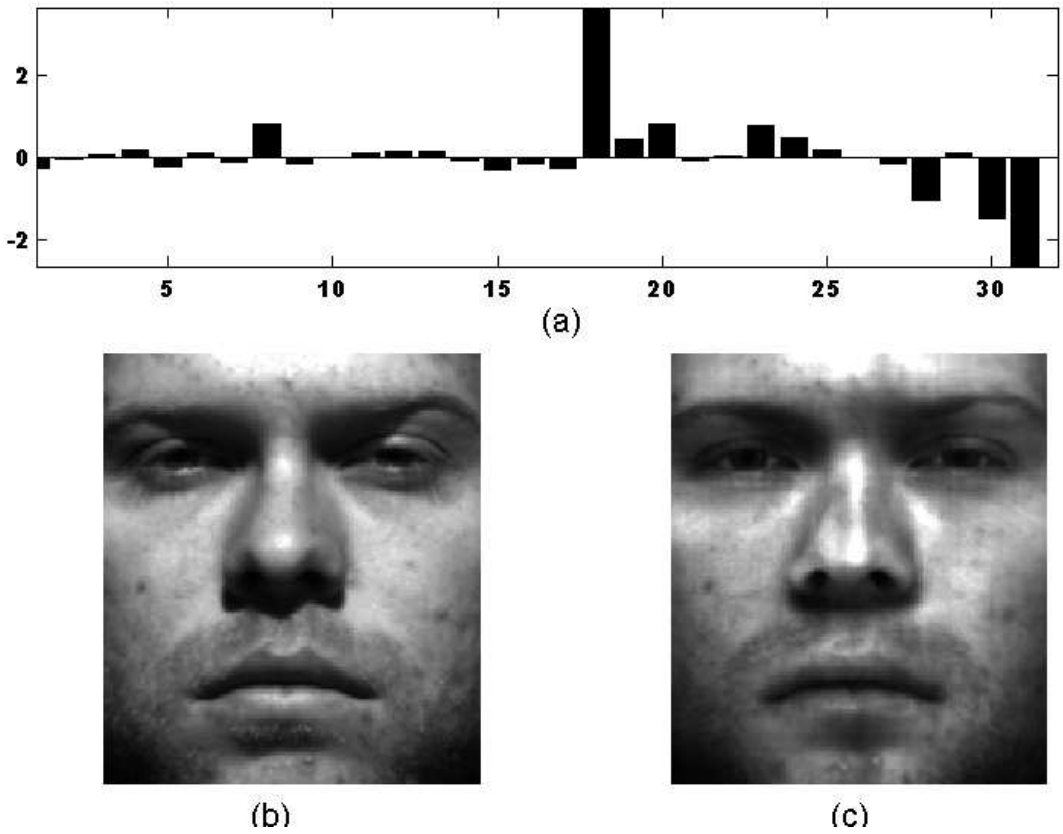


Figure 4.4: (a) the coefficient for the linear combination, (b) the input image, which is not observed in the images for training the 32 illumination conditions, and (c) the reconstructed image.

learned components can also be used for identifying which conditions the new image is associated with. These two scenarios are further evaluated in the following two subsections, with real face images.

4.3.3 Recognizing the Face Images

In this subsection, we demonstrate the performance of the proposed method in face recognition task, with the comparison to SRC, Volterraface and SUN on the extended YaleB dataset and CMU PIE dataset. As these two datasets are dominated by illumination conditions, we use the Retinex model for the proposed method, i.e., the image is converted to logarithm. In the SRC method, we build the dictionary by containing all the training images as its columns. Since there is no code publicly available for SRC, we build our own implementation. For ℓ_2 optimization used by

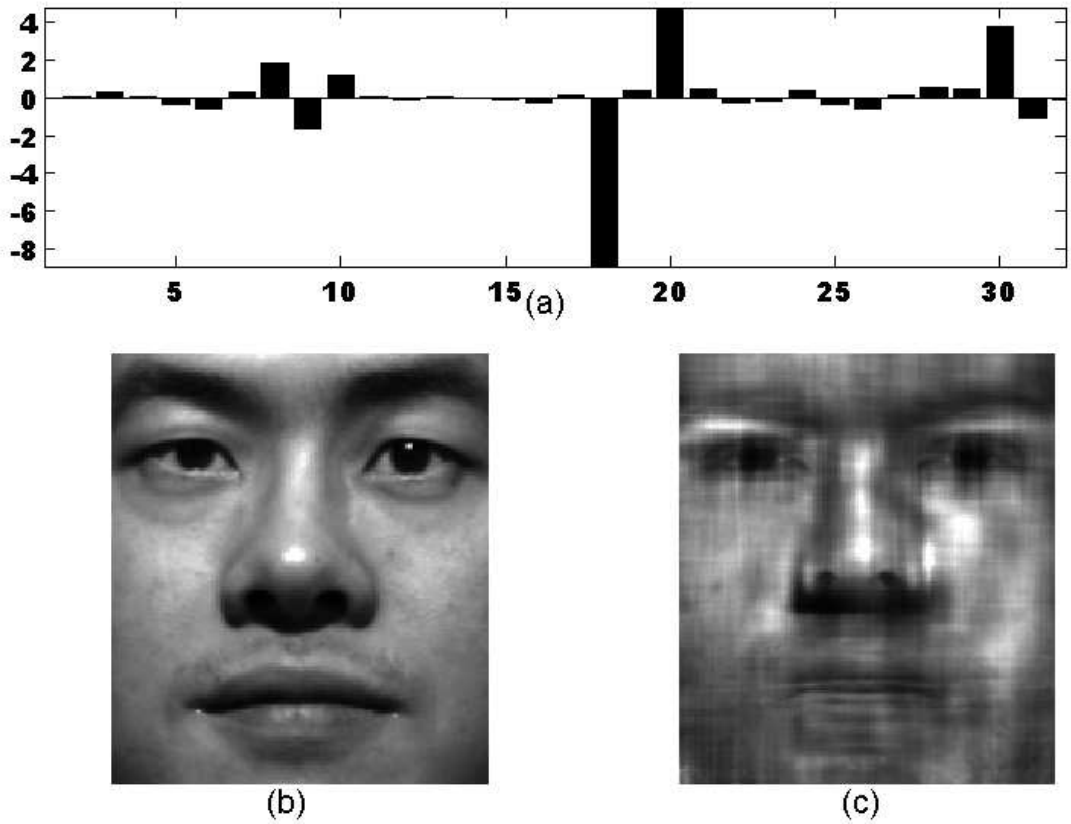


Figure 4.5: (a) the coefficient for the linear combination, (b) the input image and (c) the reconstructed image.

SRC, we used Orthonormal Matching Pursuit (OMP)Aharon *et al.* (2005) as the solver. We set the number of non-zero elements in the sparse coefficient (refer as K later) to be twice the number of conditions in the training data. In addition, each image is normalized to have zero mean and unit l_2 norm for SRC. For Volterrafaces and SUN, we use the author's original implementation and the provided parameters. For all the results, we present the both mean and standard deviation of the accuracy of 3 rounds of experiments.

To examine the robustness of the approaches with respect to the amount of training data, we use the following scheme. In the experiment, we only pick "#train per subject" images for each subject as the training instances, according to the randomly generated sample matrix (Ω), where some of the elements are set to 0 and the corresponding images won't be used for training.

The Extended YaleB dataset Georghiades *et al.* (2001) contains $N = 38$ subjects with 64 images for each subject, which correspond to 64 illumination conditions in the dataset. The images are resized to 48×42. The results on the extended YaleB dataset are summarized in Tab. 4.1. From this table, we find that the proposed approach and Volterrafaces achieve the best results; and SUN gets obviously the lowest accuracy. The performance of SRC degrades dramatically as the size of dictionary (i.e., number of training instances) reduced.

The CMU PIE dataset Sim *et al.* (2002) contains $N = 68$ subjects with varying poses, illuminations and expressions etc.. For all the images, we manually crop the face region, according to the eye position, then resize them to 50×35. The results are summarized in Tab. 4.2. In Experiment 1, all 4 methods get similar results; in Experiment 2, the proposed method and Volterrafaces get the best result; and in Experiment 3, the proposed approach gets the best result. In addition, the proposed method is more robust to the missing of training images. The performance of SRC degrades obviously as the size of dictionary reduced.

To illustrate the speed performance of the proposed approach, we compared the time required to classify one image in our approach and the SRC approach. This time was about 0.84 seconds in our method, and about 1.59 seconds in SRC. The time for the decomposition (i.e., Algorithm 1) is less than 5 minutes. The most time consuming part for the proposed approach is the singular value decomposition (SVD), which is used in computing the principle angle, so an efficient implementation of SVD can make the proposed algorithm even faster.

4.3.4 Identifying the Conditions

Finally, we use an experiment to show how the proposed method can be applied to identifying the conditions the testing images are associated with. The AR dataset Martinez and Benavente (1998) contains $N = 100$ subjects and 26 images for each subjects. The dataset contains 2 sessions, which are taken at different times. Each

(a) Experiment 1

#train per subject	32	24	16	8
Proposed	99.78±0.24%	99.54±0.04%	99.18±0.14%	**95.15±1.03%**
SRC	96.48±0.44%	95.29±0.52%	91.90±0.94%	78.65±1.81%
Volterrafaces	**99.95±0.06%**	**99.80±0.26%**	**99.48±0.49%**	90.22±11.84%
SUN	89.61±1.85%	87.64±2.80%	76.91±3.71%	60.17±2.09%

(b) Experiment 2

#train per subject	16	12	8	4
Proposed	**99.56±0.00%**	**99.33±0.23%**	**98.32±0.03%**	80.03±2.17%
SRC	89.14±0.00%	87.88±0.44%	81.02±0.13%	58.54±1.26%
Volterrafaces	99.25±0.34%	99.17±0.39%	96.27±4.03%	**91.03±2.43%**
SUN	79.22±0.00%	76.75±0.00%	68.86±0.00%	51.60±0.00%

Table 4.1: The results on extended YaleB dataset. Experiment 1: we randomly pick $M = 32$ illumination conditions for training and the remaining for testing, i.e., we will obtain $N = 38$ common components and $M = 32$ conditions by the proposed method. Experiment 2: we manually pick $M = 16$ illumination conditions for training and the remaining for testing.

session contains 13 conditions: 4 for expressions, 3 for illuminations, 3 for sun glasses and 3 for scarves. In our experiments, we use one session for training and the other session for testing. The images are converted to gray scale and resized to 55×40. To recognize the associated condition, we slightly change the formulation of the subspace:

$$\mathbf{S}_i = \{\mathbf{x} | \mathbf{x} = \sum_j w_j \times (\mathbf{a}_i + \mathbf{c}_j) \, \forall \mathbf{w} \in \mathbb{R}^N\} \tag{4.14}$$

$$\mathbf{S}_y = \{\mathbf{x} | \mathbf{x} = \sum_j w_j \times (\mathbf{y} + \mathbf{c}_j) \, \forall \mathbf{w} \in \mathbb{R}^N\} \tag{4.15}$$

where $\mathbf{S}_i$ is the subspace for condition i and $\mathbf{S}_y$ the subspace for the test image. The other settings were the same as those of previous face recognition experiments.

The proposed method achieves an accuracy of 91.77% in recognizing the conditions, with the confusion matrix given in Fig. 4.6, where we achieved over 96% accuracy for all but conditions 1, 2, 3 (3 expressions) and 12. This experiment again

(a) Experiment 1

#train per subject	20	15	10	5
Proposed	**100±0.00%**	**100±0.00%**	99.65±0.37%	97.49±0.21%
SRC	99.88±0.07%	99.88±0.07%	99.73±0.14%	**97.73±0.54%**
Volterrafaces	**100±0.00%**	**100±0.00%**	**100±0.00%**	95.83±4.16%
SUN	**100%**	99.84±0.11%	99.45±0.43%	95.75±0.49%

(b) Experiment 2

#train per subject	12	9	6	3
Proposed	**100±0.00%**	99.960.08%	99.17±0.15%	**94.70±0.20%**
SRC	99.91±0.16%	98.89±1.74%	96.90±3.73%	87.18±1.78%
Volterrafaces	**100±0.00%**	**100±0.00%**	**99.54±0.31%**	94.30±4.72%
SUN	**100±0.00%**	99.84±0.05%	98.53±0.29%	88.75±4.72%

(c) Experiment 3

#train per subject	40	30	20	10
Proposed	**99.98±0.03%**	**99.92±0.06%**	**99.24±0.06%**	**90.95±0.70%**
SRC	**99.98±0.03%**	99.45±0.03%	96.79±0.28%	86.98±0.16%
Volterrafaces	99.60±0.22%	98.37±0.47%	97.63±0.28%	89.72±1.45%
SUN	99.93±0.05%	99.38±0.14%	97.89±0.30%	88.29±0.02%

Table 4.2: The result on CMU-PIE dataset. Experiment 1: we pick the images with frontal pose (C27), which include 43 illumination conditions for each subject. We randomly pick $M = 20$ conditions for training and the remaining for testing. Experiment 2: we again only pick the image with frontal pose, but we randomly pick $M = 12$ conditions for training and the remaining for testing. Experiment 3: we use all the images from 5 near frontal poses (C05, C07, C09, C27, C29), which includes 153 conditions for each subject. We randomly pick $M = 40$ conditions for training and the remaining for testing..

demonstrates the effectiveness of the proposed method in capturing the physical conditions in the form of low-rank components.

4.4 Conclusions and Future Work

In this paper, we proposed a novel decomposition of a set of face images of multiple subjects, each with multiple images. The decomposition finds a common image and

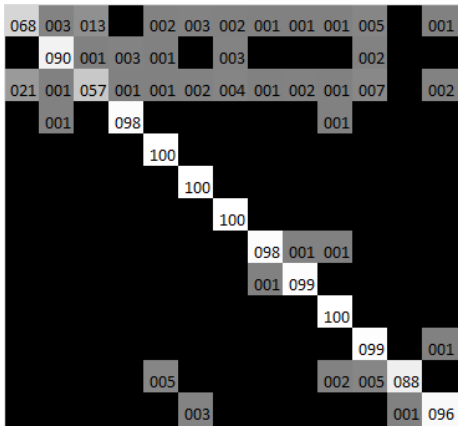


Figure 4.6: The confusion matrix (in percentage) of condition recognition result from the proposed method, where both axes are the condition index. The axis is index of the conditions.

a low-rank image for each of the subjects in the set. All the low-rank images form a set that is used to capture all possible global conditions existing in the set of images. This facilitates explicit modeling of typical challenges in face recognition, such as illumination conditions and large occlusion. Based on the decomposition, a face classifier was designed, using the decomposed components for subspace reconstruction and comparison. The classification performance shows that the proposed approach can achieve state-of-the-art performance. Experiments also showed that the proposed method is robust with missing training images, which can be an important factor to consider in a practical system. We also demonstrated with experiments that the decomposition indeed captures physically meaningful conditions, with both synthetic data and real data.

There are a few possible directions for further development of the work. In particular, the current algorithm assumes that the low-rank conditions of the training images are known and given for each of them. In practice, if the data do not have such image-level label (but still with a finite set of low-rank conditions), it is possible to expand the current algorithm by incorporating another step that attempts to estimate a mapping matrix for assigning a condition label to each image, during the

optimization iteration. This problem can be formulated as following:

$$\begin{aligned} \mathbb{C}, \mathbb{A}, \mathbb{E} \quad &= \quad \underset{\mathbb{C},\mathbb{A},\mathbb{E}}{\operatorname{argmin}} \sum_{i,j} \|\mathbf{A}_j\|_* + \lambda_{i,j} \|\mathbf{E}_{i,j}\|_1 \\ &s.t. \quad \mathbf{X}_{i,j} = \mathbf{C}_i + \sum_k \mathbf{A}_k \alpha_{i,j}(k) + \mathbf{E}_{i,j}, \ \forall \mathbf{X}_{i,j} \in \mathbb{X} \\ &\quad \|\alpha_{i,j}\|_1 = 1, \ \|\alpha_{i,j}\|_0 = 1, \ \forall (i,j) \in \Omega \end{aligned} \tag{4.16}$$

where $\alpha_{i,j}$ is a vector indicating the condition of $\mathbf{X}_{i,j}$, i.e., $\mathbf{X}_{i,j}$ takes Condition k, if and only if $\alpha_{i,j}(k) \neq 0$. In Eqn. 4.16, we assume each image takes one and only one condition, which may be too restrictive in some scenarios. Thus we could also consider remove this constraint and get a relaxed problem:

$$\begin{aligned} \mathbb{C}, \mathbb{A}, \mathbb{E} \quad &= \quad \underset{\mathbb{C},\mathbb{A},\mathbb{E}}{\operatorname{argmin}} \sum_{i,j} \|\mathbf{A}_j\|_* + \lambda_{i,j} \|\mathbf{E}_{i,j}\|_1 \\ &s.t. \quad \mathbf{X}_{i,j} = \mathbf{C}_i + \sum_k \mathbf{A}_k \alpha_{i,j}(k) + \mathbf{E}_{i,j}, \ \forall \mathbf{X}_{i,j} \in \mathbb{X} \\ &\quad \|\alpha_{i,j}\|_1 \leq \tau, \ \forall (i,j) \in \Omega \end{aligned} \tag{4.17}$$

Note, we still impose a sparsity constraint on $\alpha_{i,j}$ to make the problem more determined.

Chapter 5

UNSUPERVISED VIDEO ANALYSIS BASED ON A SPATIOTEMPORAL SALIENCY DETECTOR

The third problem is related to visual saliency detection and its potential application in varying visual tasks. Visual saliency, which predicts regions in the field of view that draw the most visual attention, has attracted a lot of interest from researchers. It has already been used in several vision tasks, e.g., image classification, object detection, foreground segmentation. Recently, the spectrum analysis based visual saliency approach has attracted a lot of interest due to its simplicity and good performance, where the phase information of the image is used to construct the saliency map. In this chapter, we propose a new approach for detecting spatiotemporal visual saliency based on the phase spectrum of the videos, which is easy to implement and computationally efficient. The method is based on the prior information that, the salient objects is typically small compared with the whole spatiotemporal volume and the background, either static or dynamic, usually has sparse support in frequency domain. With the proposed algorithm, we also study how the spatiotemporal saliency can be used in two important vision task, abnormality detection and spatiotemporal interest point detection. The proposed algorithm is evaluated on several commonly used datasets with comparison to the state-of-art methods from the literature. The experiments demonstrate the effectiveness of the proposed approach to spatiotemporal visual saliency detection and its application to the above vision tasks.

5.1 Introduction

In the recent years modeling and detection of visual saliency has attracted a lot of interest in the vision community. One early work that is widely known is the approach by Itti *et al.* Itti *et al.* (1998). Since then, a lot of different models have been proposed

for computing visual saliency. Such models may be roughly divided into two groups: bottom-up models (or stimulus driven) that are mainly based on low-level visual features of the scene, and top-down model (goal-driven) that employs information and knowledge about a visual task. A survey of both groups of methods was reported in Borji and Itti (2012). Visual saliency analysis has been applied with success to other vision tasks including object detection Alexe *et al.* (2012), image classification Sharma *et al.* (2012), foreground segmentation Li *et al.* (2011) and securities Zhao *et al.* (2013).

Recently, spectral-based approach has gained increased interest due to its simplicity and good performance. In Hou and Zhang (2007), the spectrum residual together with the phase information was used to construct a saliency map. In Guo *et al.* (2008) it was found that it is the phase information rather than the spectrum leads to a better saliency map. However, there was a lack of theoretic justification for such methods until Hou *et al.* (2012), where it was shown that, if the background is sparsely supported in the DCT domain and the foreground is sparsely supported in the spatial domain the foreground will receive high value on the computed saliency map.

In the real world, the visual field-of-view of a human may constantly change, and thus visual saliency often depend on not only a static scene but also the changes in the scene. To this end, spatiotemporal saliency has been proposed, which tries to capture regions attracting visual attention in the spatiotemporal domain. Spatiotemporal saliency has been applied to vision tasks such as video summarization Ma *et al.* (2005), human-computer interaction Itti *et al.* (2004), video compression Guo and Zhang (2010), and abnormality detection Gao *et al.* (2009).

In this chapter we propose a novel spatiotemporal visual saliency detector for video analysis, based on *the phase information of the video.* With the saliency map computed using the proposed method, we analysis how it can be used for two fundamental vision tasks, namely abnormality detection and spatiotemporal interest point detec-

tion. We evaluate the performance of the proposed algorithm using several widely used datasets, with the comparison to the state-of-art in the literature.

The proposed method, compared with the existing work on spatiotemporal saliency in the literature, has several advantages. First, it computes the saliency information from the entire video span, which is different from many existing approaches in the literature. For example, Guo *et al.* (2008) computes temporal information by only the differences of two adjacent frames, which is insufficient for modeling complex activities, as shown in the experiment part. Second, the proposed approach is easy to implement and computationally efficient. The core parts of the algorithm involve only a three-dimensional Fourier transform, whose complexity is only $O(N \log N)$, where N is the size of the input. Last but not least, no training stage or prior information is needed for the proposed approach, which is a significant advantage for applications like abnormality detection.

The rest of the chapter is organized as follows: in Sec. 5.2 we describe the proposed method including the analysis and the relationships with the existing methods; Sec. 5.3 evaluate the proposed spatiotemporal saliency detector in saliency detection on both synthetic dataset and two real video dataset; studies of how the spatiotemporal saliency computed by the proposed method can be used for two important vision tasks, abnormality detection and spatiotemporal interest point detection, is presented in Sec. 5.4; and the chapter is concluded in Sec. 5.5.

5.2 Proposed Method

As reviewed above, spectrum analysis based approaches to visual saliency has seen some success, although the existing work has been primarily on predicting salient objects for a given (static) image. For a dynamic scene, temporal information should be taken into consideration for properly predicting the salient objects. For example, it was found in Ölveczky *et al.* (2003) that objects attract more visual attention if they move differently than their neighbors. Considering this, we propose to compute

the saliency map of dynamic scenes by utilizing the phase information of the temporal domain together with the phase information of the spatial domain. In the proposed method, we compute the saliency map for 3D data $\mathbf{X} \in \mathbb{R}^{M \times N \times T}$ as:

$$\mathbf{Z} = \left| \mathcal{F}^{-1} \left(\frac{\mathbf{Y}}{|\mathbf{Y}|} \right) \right|^2 \tag{5.1}$$

where $\mathbf{Y} = \mathcal{F}(\mathbf{X})$, $\mathcal{F}$ is 3D discrete Fourier transform and $\mathcal{F}^{-1}$ is the corresponding inverse transform. After we get the saliency map, we smooth it with a 3D Gaussian smooth filter. The 3D Fourier transform can be computed as:

$$\begin{aligned} & \mathbf{Y}(u,v,w) \qquad (5.2) \\ &= \sum_t \sum_i \sum_j \mathbf{X}(i,j,t) e^{-i2\pi\left(\frac{ui}{M}+\frac{vj}{N}+\frac{wt}{T}\right)} \\ &= \sum_t e^{-i2\pi\frac{wt}{T}} \sum_i \sum_j \mathbf{X}(i,j,t) e^{-i2\pi\left(\frac{ui}{M}+\frac{vj}{N}\right)} \\ &= \sum_t e^{-i2\pi\frac{wt}{T}} \sum_i e^{-i2\pi\frac{ui}{M}} \sum_j \mathbf{X}(i,j,t) e^{-i2\pi\frac{ui}{M}} \end{aligned}$$

i.e., the 3D Fourier transform can be computed as a sequence of 1D Fourier transforms on each coordinate.

The proposed method detects spatiotemporal saliency, which has been also discussed in some existing works. For example, in Guo *et al.* (2008), the detection was done by combining color information of one frame and the differences of this frame to the previous one with quaternion Fourier transform. As a result, the temporal information is limited to two adjacent frames and is insufficient for modeling complex scenes. On the other hand, the spatiotemporal saliency proposed in this chapter considers the temporal information over a much larger temporal span, which is up to the entire video.

The method in Eqn. 5.1 evaluates the saliency of a region by exploring the information of the entire video. In some situations, we may also be interested in detecting a region that is salient within a temporal window of the video. For example, if a video contains multiple scenes, each capturing a different activity, we may be more interested in analysis the saliency within each scene instead of the entire video. For this

reason, we propose multi-scale analysis for spatiotemporal saliency, which is inspired by short-time Fourier transform. We first apply the window function to the input signal $\mathbf{X} \cdot \mathbf{w}(i, j, t)$, where $\cdot$ is the element-wise multiplication and $\mathbf{w}(i, j, t)$ the window function centered at position (i, j, t), which is nonzero for only a small support (i.e., the size of window function). The saliency map is computed for the windowed signal:

$$\begin{aligned} \mathbf{Y} &= \mathcal{F}[\mathbf{X} \cdot \mathbf{w}(i, j, t)] \\ \mathbf{Z}(i, j, t) &= \mathcal{F}^{-1}\left[\frac{\mathbf{Y}}{|\mathbf{Y}|}\right] \end{aligned} \tag{5.3}$$

By sliding the window function on the input video, we still obtain the saliency map for the entire video. The size of the sliding window determines the temporal resolution: with a larger window, more global information of the input is revealed but the resolution is lower; with a smaller window, resolution is improved. The window function can be applied in either temporal direction, spatial direction or both. As a result, we can perform saliency detection from varying scales, which enables us to reveal the information at different spatiotemporal resolution, similar to short time Fourier transform.

Combining different visual cues is important for not only scene saliency but also spatiotemporal saliency. In this chapter, we proposed to compute the saliency map for each cue independently then compute the summation of saliency maps from all visual cues. In Guo and Zhang (2010), quaternion Fourier transform (QFT) is utilized to combine the three-channel color information and frame differences. However, the QFT could be very expensive (e.g., time consuming) when applied in spatiotemporal domain. In fact we find that (Appendix B): given a data with four feature channels, the saliency map computed with QFT is very similar with the sum of saliency maps computed with FFT over each feature channel.

Finally, we summarize the proposed algorithm below:

Algorithm for Spatio-Temporal Visual Saliency Detection

Input: data $\mathbf{X}$, Gaussian filter g, window function $\mathbf{w}$

Output: saliency map $\mathbf{Z}$

for each window location **do**

  for each feature channel **do**

    Apply $\mathbf{w}$ to the input $\mathbf{X}$;

    Compute Fourier coefficient $\mathbf{Y} = \mathcal{F}[\mathbf{X}]$;

    Extract the phase information $\hat{\mathbf{Y}} = \frac{\mathbf{Y}}{|\mathbf{Y}|}$;

    Do the inverse transform $\mathbf{Z} = \left|\mathcal{F}^{-1}[\hat{\mathbf{Y}}]\right|^2$;

    Smooth saliency map $\mathbf{Z} = \mathbf{Z} * g$;

  end for

  Combine the $\mathbf{Z}$ of all channels together;

end for

where $\mathbf{W}$ is the window function. Currently, we only apply the window function along temporal direction and rectangle window is used. The size of the window is depending on the data. By incorporating the phase information of the temporal domain, the proposed method can not only suppress the static background, as achieved by visual saliency for images, but also suppress the object which is static or moving "regularly" as will be presented in Sec: 5.3.1.

5.2.1 Analysis

There has been several explanations for why spectral domain based approach is able to detect saliency region from the image. For example, Bian and Zhang (2009) explained by its biological plausibility that saliency map exists in the primary visual cortex (V1), which is orientation selective and lateral surround inhibition Simoncelli and Schwartz (1999). The spectral magnitude measures the total response of cells

tuned to the specific frequency and orientation. According to lateral surround inhibition, similarly tuned cells will be suppressed depending on their total response, which can be modeled by dividing its spectral by the spectral magnitude Zhaoping and Dayan (2006). Hou *et al.* (2012) provided another explanation from sparse representation which states that, if the foreground is sparse in spatial domain and background is sparse in DCT domain (e.g., periodic textures), the spectral domain based approach will highlight the foreground region in the saliency map.

Motion, like color and texture, is also perceptually salient. Huber and Healey (2005) studied how three properties of motion, namely flicker, direction and velocity, contribute to this saliency. By setting the target object having different flicker rate, moving direction or motion velocity from the other objects, the target object can be easily identified by human subjects, i.e., being salient. In spectral, the target object and other objects can be mapped to two different bands (frequency and orientation), where the band corresponding to the target object has a much lower response than the band for the other object. Thus if we set the magnitude of the spectral to one, as the proposed method dose, the band for the other objects will be suppressed more than the target object, which makes the target object "popped out" in the output. In Sec. 5.3.1, we will verify this analysis with experiments on synthetic data.

5.2.2 Relationship to Existing Works

Our method is related to some existing works and based on the way in which they computed the temporal information, we can roughly divide them into two categories:

1. Methods of the first category represent the temporal information by the motion, e.g., by frame differences Ma *et al.* (2005) or by more dedicated motion estimation method including homography of adjacent frames Zhai and Shah (2006) and phase correlation Bian and Zhang (2009). However, methods of this temporal information typically have limited temporal span, e.g., two adjacent frames (Zhang *et al.* (2009) tried to compute the frame differences of frames at

a predefined sets of temporal spans), thus they are not sufficient for modeling the complex motion patterns.

2. In this category, the saliency of a spatiotemporal cuboid (refer as cuboid later) is measured by the "differences" of this cuboid to other cuboids of the video or the template in the dictionaries, which may require high computational cost and/or require additional training data. The "differences" of cuboids can be measured by distances Seo and Milanfar (2009), relative entropy Li *et al.* (2010) mutual information Mahadevan and Vasconcelos (2010) and coding length increments Ban *et al.* (2008).

The proposed method is different from these methods. First, it does not rely on prior knowledge. Instead, it explores within the input video to detect the potential "outliers". Second, the "outliers" are found by exploring the whole temporal span. This makes the proposed algorithm be able to detect salient patterns from complex dynamic background. In addition, the propose method has low computational costs and is easy to implement. Fourier transform for multiple dimensional data can be computed as a sequence of 1D Fourier transform on each coordinate of the data, thus the computational cost of 3D Fourier transform for data $X \in \mathbb{R}^{M\times N\times T}$ is $O(MNT\log(MNT))$. Thus the total computation cost for the proposed algorithm is $O(KMNT\log(MNT))$, where K is the number of feature channels.

5.3 Experiment

In this section, we evaluate the proposed method in saliency detection on both syntheic data (Sec. 5.3.1) and on two real image datasets (Sec. 5.3.2), CRCNS-ORIG and DIEM. The performances of the proposed methods are compared with the existing methods, some of which are state-of-art.

5.3.1 Simulation Experiment

In this section, we evaluate the proposed method on synthetic data. In Huber and Healey (2005), how three properties of motion, namely flicker, direction and velocity, contribute to the saliency was studied. In this section, we generate the synthetic data according to their protocol. The input data is a short clip where the resolution is 174×174 with 400 frames at the frame rate of 60 frames per second. We put 36 objects of size 5×13 in a 6×6 grid and a target object is randomly selected out of those 36 objects. All the objects are allowed to move within a 29×29 region centered at their initial position (and warped back, if they move out of this region). The video is black-and-white. We design the following three experiments:

1. **Flicker**: we set the objects on-off at a specified rate and the target object at a different rate from the other 35 objects;

2. **Direction**: we set the objects moving in a specified direction and the target object in a different direction. The velocities of all the objects are the same;

3. **Velocity**: we set the objects moving in a specified velocity and the target object moves in a different velocity. The moving direction of the all the objects are the same.

All the other parameters are the same as used in Huber and Healey (2005). According to Huber and Healey (2005), the target object could be easily identified by human subjects, when its motion property (e.g., flicker rate, moving direction and velocity) is different from the other objects. We also include some "blind" trials, where the target object has the same motion property as the other 35 objects. In this case, the target object can't be identified by the human subjects, i.e., there is no salient region.

We apply the proposed method to the input data. For comparison, we also evaluate the method proposed in Bian and Zhang (2009) and Hou *et al.* (2012). We use the area under receiver operating characteristic curve as the performance metric. The

ground truth mask is generated according to the location of the target object. The experiment result is shown in 5.1.

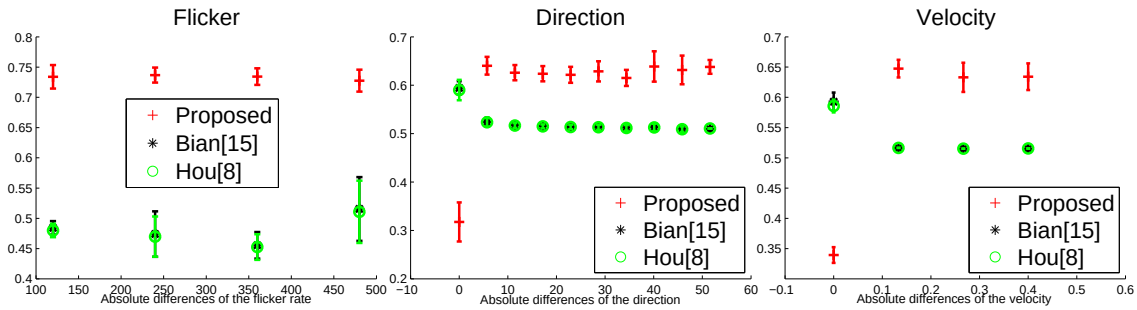
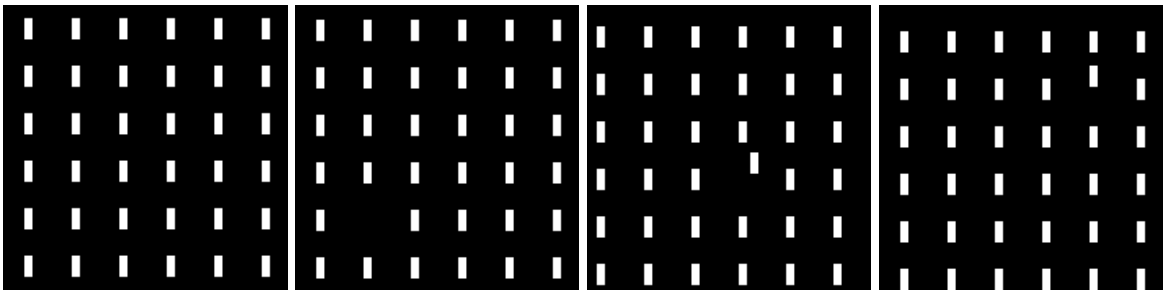


Figure 5.1: The AUC on the synthetic data for the proposed method and two existing methods. For "Direction" and "Velocity", we also include some "blind" trials (X-axis has value 0), where the target object has exactly the same motion property as the other 35 objects. In those trials, the target object can't be identified by human subjects, i.e., there is no salient object Huber and Healey (2005).



Blind Flicker Direction Velocity

Figure 5.2: Some visual sample of the synthetic data for different experiments.

From the experiment results, we can find that the proposed method detects the salient region much more accurately than Bian and Zhang (2009) and Hou *et al.* (2012) in all except the "blind" trials. For the "blind" trials, the AUC for the proposed method significantly reduces, which shows that the proposed method is also robust. However, Bian and Zhang (2009) and Hou *et al.* (2012) don't survive in those "blind" trials. Surprisingly, Bian and Zhang (2009) and Hou *et al.* (2012) achieves quite similar performances, though Bian and Zhang (2009) was supposed to achieve better result as it include the differences of two adjacent frames as motion (temporal) information.

5.3.2 Spatiotemporal Saliency Detection

In previous section, we test the proposed spatiotemporal saliency detector on synthetic videos, with the comparison to two other saliency detectors, where the proposed detector shows better performances in capturing the temporal information. In this section, we evaluate the proposed spatiotemporal saliency detector on two challenging video datasets for saliency evaluation, CRCNS-ORIG Itti (2009) and DIEM Mital *et al.* (2011). For this experiment, we first convert each frame into the LAB color space, then compute the spatiotemporal saliency in each channel independently and the final spatiotemporal saliency is the summation of the saliency maps of all three channels.

CRCNS-ORIG includes 50 video clips from different genres, including TV programs, outdoor scenes and video games. Each clip is 6-second to 90-second long at 30 frames per second. The eye fixation data is captured from eight subjects with normal or correct-normal vision. In our experiment, we down-sample the video from 640×480 to 160×120 and keep the frame rate untouched, then apply our spatiotemporal saliency detector. To measure the performance, we compute the area under curve (AUC) and F-measure (harmonic mean of true positive rate and false positive rate). The experiment result is shown in Fig. 5.3, where the area under curve (AUC) is 0.6639 and F-measure is 0.1926. Tab. 5.1 compares the result of the proposed method with some state-of-art methods on CRCNS-ORIG, which indicates that our method outperforms them by at least 0.06 regarding AUC. The per-video AUC score is shown in Fig. 5.4.

DIEM dataset collects data of where people look during dynamic scene viewing such as film trailers, music videos, or advertisements. It currently consists of data from over 250 participants watching 85 different videos. Each video in DIEM dataset includes 1000 to 6000 frames at 30 frames per second. Similarly as CRCNS, we down-sample the video to 1/4 (e.g., from 1280×720 to 320×180) while maintaining the aspect ratio and frame rate. We observe that each video in DIEM dataset is consisted

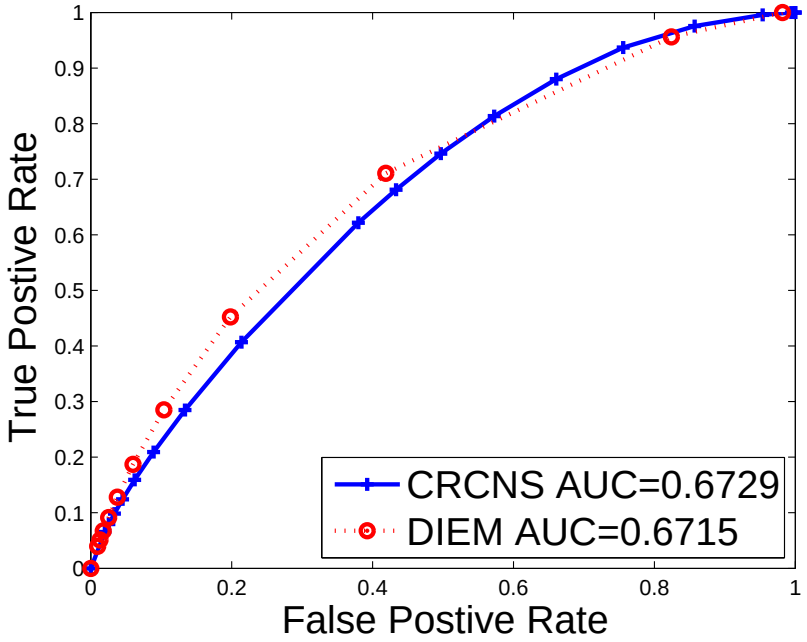


Figure 5.3: The receiver operating characteristic curve of the propose method in CRCNS-ORIG dataset and DIEM dataset. The area under the curve is 0.6639 and 0.6896 accordingly.

of a sequence of short clips, where each clip has 30 to 100 frames. To properly detect the saliency from those videos, we apply the window function to our spatiotemporal saliency detector, where the size of the window (along temporal direction) is 60-frame. The experiment result is shown in Fig. 5.3 and Tab. 5.1, where the AUC is 0.6896 and F-measure is 0.35. From the table, we can find that the proposed method outperforms the state-of-arts by over 10%. The per-video AUC score is shown in Fig. 5.5.

5.4 Application of Spatiotemporal Saliency

In the previous section, we show that the proposed method is able to detect the saliency region in the video. The saliency detection for image has been used more and more in other visual tasks, e.g., image segmentation, object recognition. A natural question arises that can we also apply the spatiotemporal saliency detection for some important vision tasks. In this section, we show how can we applied the spatiotemporal saliency computed by the proposed methods to two important vision

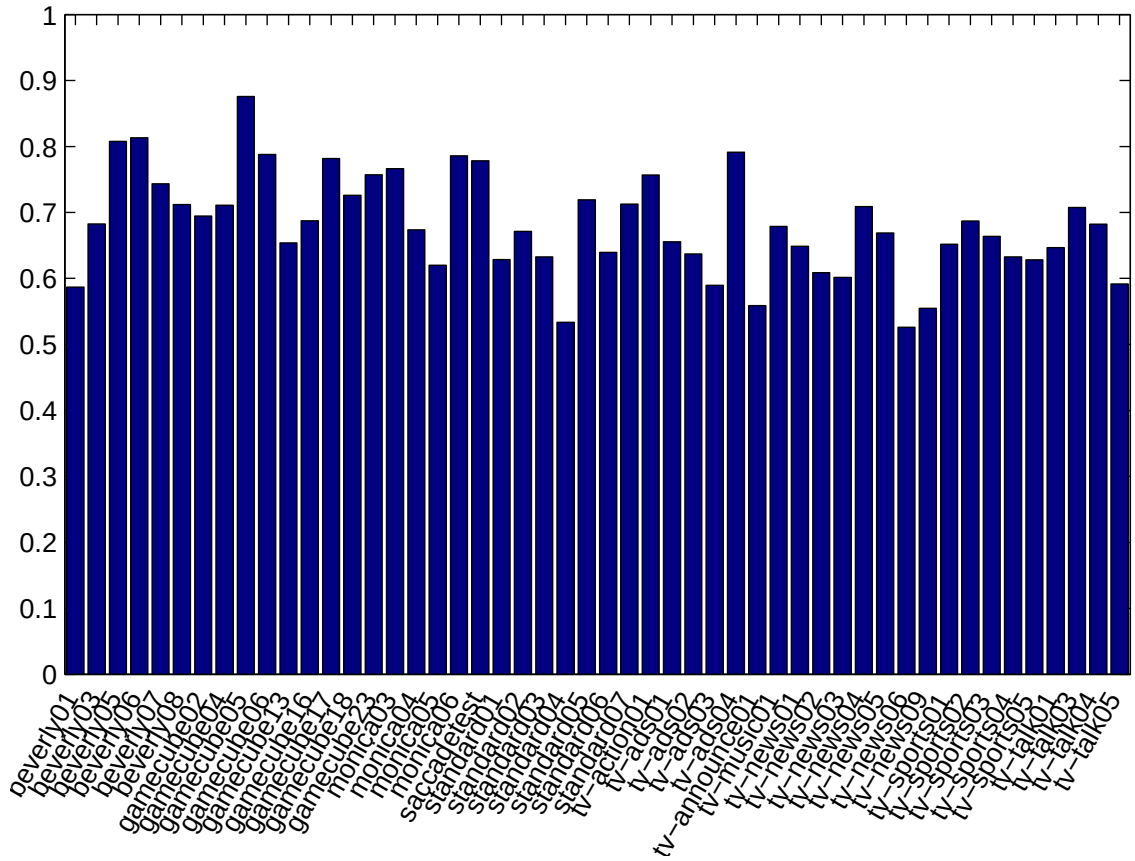


Figure 5.4: The AUC of the proposed method for each video from CRCNS-ORIG dataset.

tasks, i.e., abnormality detection (5.4.1) and spatiotemporal interest pointer detection (5.4.2).

5.4.1 Abnormality Detection

According to our previous analysis, the salient region should be different from the neighbor, both spatially and temporally. This spatiotemporal saliency shares a lot of common to the concept of abnormality in video. Thus in this section, we show how can we utilize the proposed spatiotemporal saliency detector to detect abnormality from the video.

For abnormality detection, we start with computing the saliency map for the input video as described above. The regions containing abnormalities can be detected by founding the region where the saliency value is above a threshold. Then the saliency score of a frame is computed as the average of saliency value of the pixels in that

Method	AUC	Method	AUC
AWS Garcia-Diaz *et al.* (2009)	0.6000	AWS Garcia-Diaz *et al.* (2009)	0.5770
HouNIPS Hou and Zhang (2008)	0.5967	Bian Bian and Zhang (2009)	0.5730
Bian Bian and Zhang (2009)	0.5950	Marat Marat *et al.* (2009)	0.5730
IO	0.5950	Judd Judd *et al.* (2009)	0.5700
SR Hou and Zhang (2007)	0.5867	AIM Bruce and Tsotsos (2005)	0.5680
Torralba Torralba (2003)	0.5833	HouNIPS Hou and Zhang (2008)	0.5630
Judd Judd *et al.* (2009)	0.5833	Torralba Torralba (2003)	0.5840
Marat Marat *et al.* (2009)	0.5833	GBVS Harel *et al.* (2006)	0.5620
Rarity-G Mancas (2007)	0.5767	SR Hou and Zhang (2007)	0.5610
CIOFM Itti and Baldi (2006)	0.5767	CIO Itti and Baldi (2006)	0.5560
Proposed	0.6639	Proposed	0.6896

Table 5.1: The result the proposed method compared with the results of the top ten existing methods on CRCNS dataset (left) and DIEM dataset (right) according to Borji *et al.* (2012). From this table, we can find that the propose method gets obvious better performances than the state-of-arts on both two datasets.

frame, i.e.,

$$\mathbf{s}(t) = \frac{1}{NM} \sum_i \sum_j \mathbf{X}(i, j, t) \tag{5.4}$$

where $\mathbf{s}(t)$ is the saliency score of t_{th} frame, $N \times M$ is the size of one frame, i, j, t are row, column and frame index of the 3D saliency map accordingly. The frame with high saliency score would contain abnormality.

We evaluate the proposed method for abnormality detection in videos from two datasets: UMN abnormal dataset[1] and UCSD dataset Mahadevan *et al.* (2010). Abnormal detection has attracted a lot of efforts from the researchers. However, most of the existing works require training stage, e.g., social force Mehran *et al.* (2009), sparse reconstruction Cong *et al.* (2011), MPPCA Kim and Grauman (2009), i.e., they need training data to initialize the model. The proposed method, instead, dose **NOT** need any training stage or training data.

[1] `http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi`

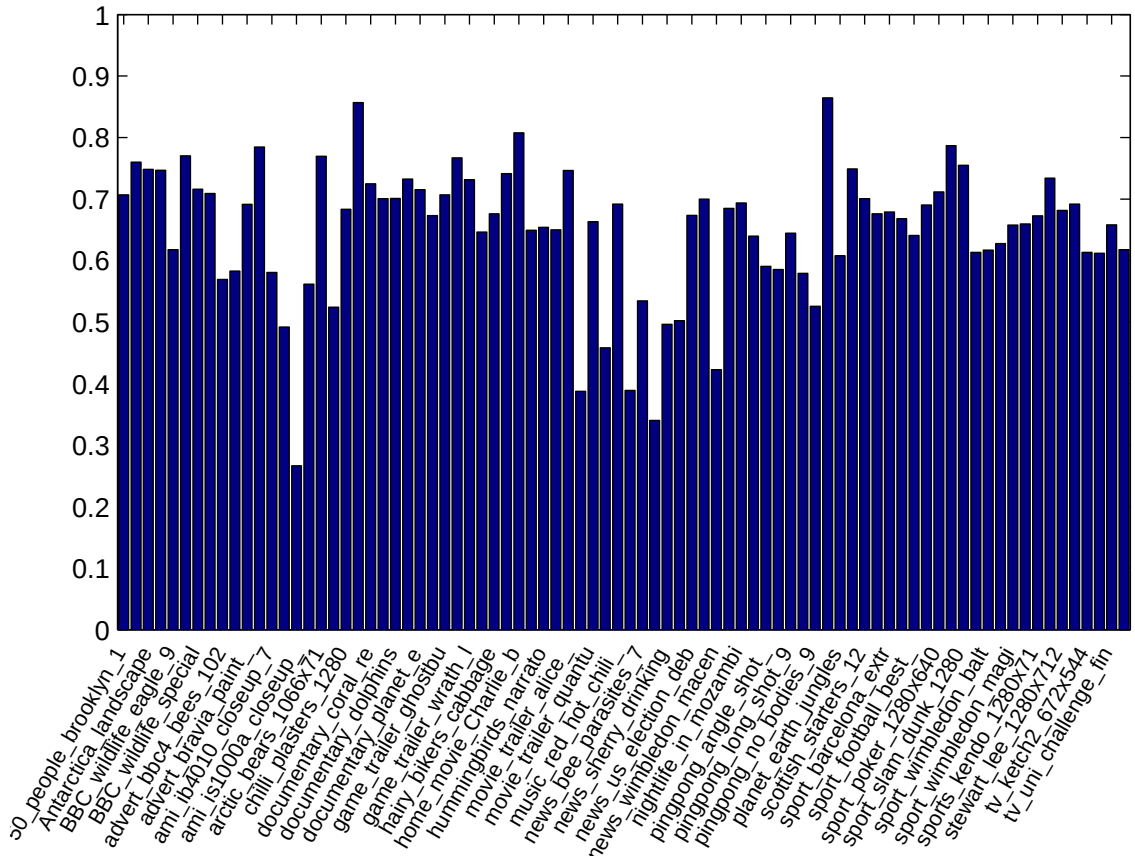


Figure 5.5: The AUC of the proposed method for each video from DIEM dataset.

The result on UMN abnormal dataset is shown in Tab. 5.2, where we compute the frame-level true positive rate and false positive rate then compute the area under the ROC (Fig. 5.6). Fig. 5.7 shows the result for videos of three scenes, where we plot saliency value of each frame and show some sample frames. The result on UCSD dataset is shown in Tab. 5.3, where we report frame-level equal-error rate (EER) Mahadevan *et al.* (2010). Fig. 5.8 shows the ROC for UCSD dataset with the proposed method; Fig. 5.9 shows eight samples frames, where red color highlights abnormal regions. We can find that, without training data, the proposed method still outperforms several state-of-arts in the literature, e.g., social force, MPPCA.

5.4.2 Spatiotemporal Saliency Point Detector

The regions which attract human's attention most would contribute most to people's perception of the scene. The saliency map computed with the proposed method

Method	AUC
Optical flow Mehran *et al.* (2009)	0.84
Social force Mehran *et al.* (2009)	0.96
Chaotic invariants Wu *et al.* (2010)	0.99
NN Cong *et al.* (2011)	0.93
Sparse reconstruction Cong *et al.* (2011)	0.978
Interaction force Raghavendra *et al.* (2011)	0.9961
Proposed	0.9378

Table 5.2: The result on UMN dataset. Note, we have cropped out the region which contains the text "abnormal", and results in frame resolution 214×320. Please note that, most of those methods, except the proposed one, need a training stage.

Method	Ped1	Ped2	Overall
Social force Mehran *et al.* (2009)	31%	42%	37%
MPPCA Kim and Grauman (2009)	40%	30%	35%
MDT Mahadevan *et al.* (2010)	25%	25%	25%
Adam Adam *et al.* (2008)	38%	42%	40%
Reddy Reddy *et al.* (2011)	22.5%	20%	21.25%
Sparse Cong *et al.* (2011)	19%	*N.A.*	*N.A.*
Proposed	27%	19%	23%

Table 5.3: The frame level EER (the lower the better) for UCSD dataset. Please note that, most of those methods, except the proposed one, need a training stage. From the result, we can found that the proposed method, even without traing stage or training data, can still outperform social force, MPPCA.

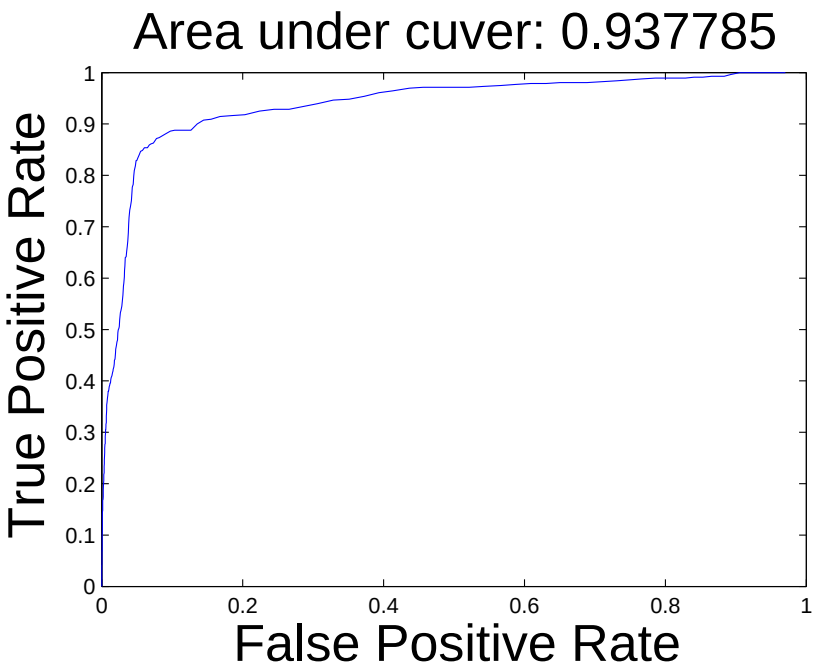


Figure 5.6: The ROC for the UMN dataset computed with the propose method.

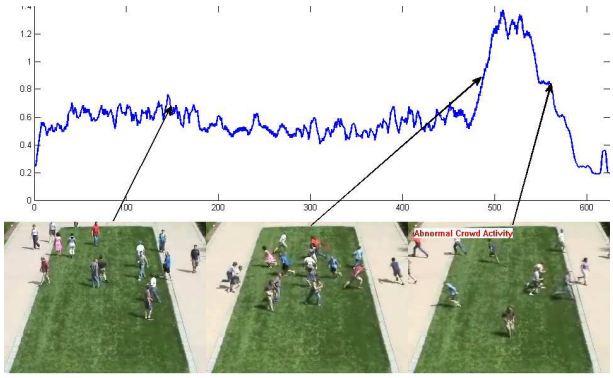
will highlight those regions. Thus we propose to sample the interest points based on the saliency map of the data, which we refer as spatiotemporal saliency point detector (STSP).

To detect interest point, we also starts with computing the saliency map $\mathbf{Z}$ for the input data $\mathbf{X}$. Then we apply non-maximum suppression on the saliency map to sample the interest points: an interest point is selected at (x, y, t) if and only if

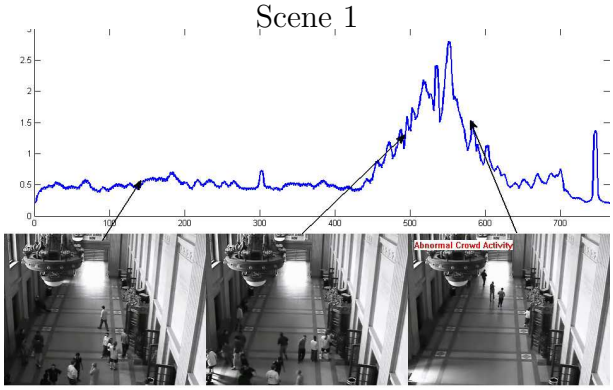
$$\begin{aligned} \mathbf{Z}(x,y,t) &\geq \rho \\ \mathbf{Z}(x,y,t) &\geq \mathbf{Z}(i,j,k)\ \forall (i,j,k) \in N(x,y,t) \end{aligned} \tag{5.5}$$

where ρ is a predefined threshold (e.g., 2μ) and $N(x, y, t)$ is the set of positions near (x, y, t).

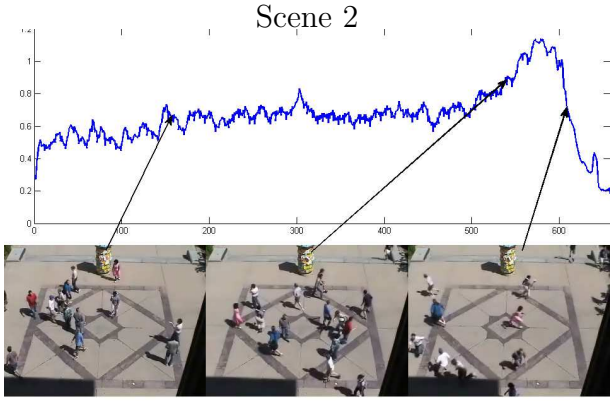
Similar as Laptev (2005), for each interest points (x, y, t), we extract a descriptor within its neighbor area characterized as (x, y, t, σ, τ), where (x, y, t) is the center, σ, τ are the spatial and temporal scales (we use $18 \times 18 \times 10$, $25 \times 25 \times 14$ and $36 \times 36 \times 20$ here) accordingly. The neighbor is further divided into multiple sub-blocks (e.g., $3 \times 3 \times 2$ along spatial and temporal direction accordingly); for each sub-block, we



Scene 1



Scene 2



Scene 3

Figure 5.7: Some sample results for the UMN datasets, where we pick one video for each scene. The top is the saliency value (Y-axis) for each frame (X-axis) and bottom are sample frames picked from different frames (as shown by the arrow).

computed the 3D gradient $g = [g_x, g_y, g_t]$; then we quantize the orientations of the gradients into a histogram of four bins; finally the histogram of each sub-block is normalized to unit l_1 norm and histograms of all sub-blocks is concatenated into one histogram, i.e., the descriptor for interest point (x, y, t).

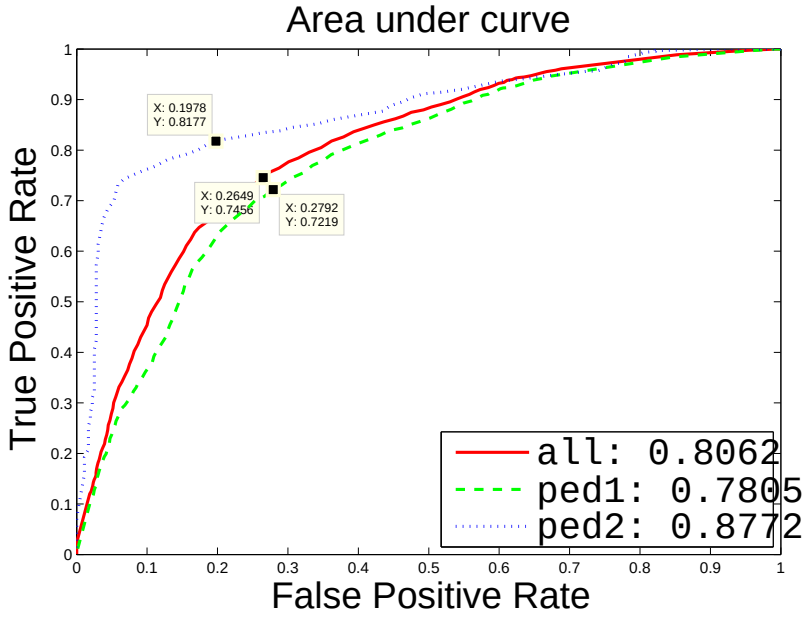
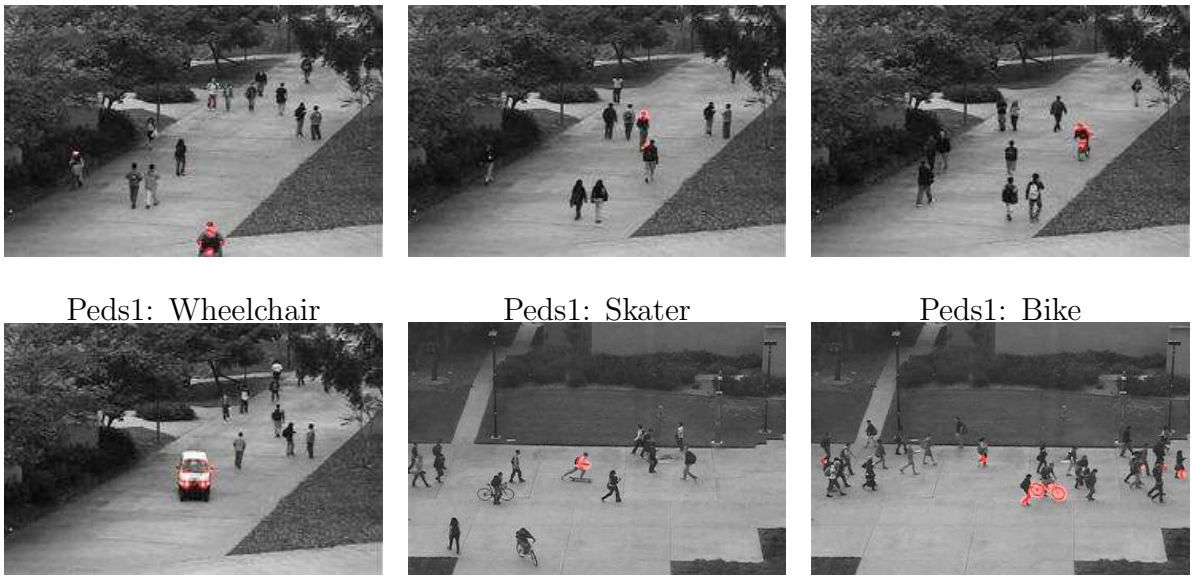


Figure 5.8: The ROC for the UCSD dataset computed with the propose method.



Peds1: Wheelchair | Peds1: Skater | Peds1: Bike

Peds1: Cart | Peds2: Skater | Peds2: Bike

Figure 5.9: Some sample results for the UCSD datasets, where the red color highlights the detected abnormal region, i.e., the saliency value of the pixel is higher than four times of the mean saliency value of the video.

Compared with existing spatiotemporal interest point detectors, which mostly choose the location where the gradient is strong and stable cross different scales. However, the gradient is low level information and is insufficient to capture the complex dynamics as the human vision does. Instead, the proposed method explores the

relationship of each location over all spatial and temporal spans, which is able to model complex dynamics in the video.

For evaluation, we use three datasets: Weizmann dataset Gorelick *et al.* (2007), KTH dataset Schuldt *et al.* (2004) and UCF sports dataset Rodriguez *et al.* (2008). Since the method is proposed for detecting interest points, we only compare it with several state-of-art spatiotemporal interest point detectors including Harris3D Laptev (2005), Gabor Dollar *et al.* (2005), Hessian3D Willems *et al.* (2008) and dense sampling Kläser (2010), where the result are summarized in Kläser (2010). The parameters of the detectors are set as suggested by the chapter accordingly.

Fig. 5.10 shows the saliency map for some sample frames of videos from UCF sports action dataset and KTH dataset. From the figure, we can found that the saliency map computed with the proposed method highlights the moving region while suppressing the background. The proposed method is also robust to moving background (Row 1), clutter background (Row 2) and scale variation (Row 3). In addition, from Row 3 to 4, we can found the moving parts of body, e.g., hands, get higher saliency value (red color) then other body parts. The spatiotemporal saliency interest points will be mostly sampled from those highlighted regions and augment the description of the action of interest.

To quantitatively evaluate the performances of different detector, we use the interest points detected by those detectors to train a classifier for activity recognition. We use both histogram of gradient (HoG) and histogram of optical flow (HoF) as the descriptor. Bag of words is used to represent the video, where each input is represented as a histogram of words in the code-book (size of of code-book is $k = 2000$); then classifier (support vector machine with χ^2 kernel) is applied to those histograms to classify the input. For Weizmann dataset and UCF sports dataset, we use leave-one-out scheme for training and testing; for KTH dataset, we follow the standard partition in Schuldt *et al.* (2004).

Method	Weizmann	KTH	UCF sports
Harris3D	85.6%	91.8%	78.1%
Gabor	N.A.	88.7%	77.7%
Hessian3D	N.A.	88.7%	79.3%
Dense	N.A.	86.1%	81.6%
Proposed	84.5%	88.0%	**86.7**%
Proposed*	**95.6**%	**92.6**%	85.6%

Table 5.4: The performances of different detectors on three datasets. For "proposed*", we extract the descriptor on the saliency map instead of on the video.

Tab. 5.4 reports the performances of different detectors on three dataset, where we test extracting feature on the original video and also extracting feature on the saliency map of the original video (refer as "proposed*"). From the table we find that, the proposed method (especially "proposed*") achieves the best result over all three datasets. Especially "proposed*" achieved the best results for KTH dataset and Weizmann data; "proposed" achieved the best results for UCF sports action dataset. For video with simple background(e.g., KTH dataset and Weizmann dataset), extracting descriptor on saliency map instead of the video itself could be a better choice.

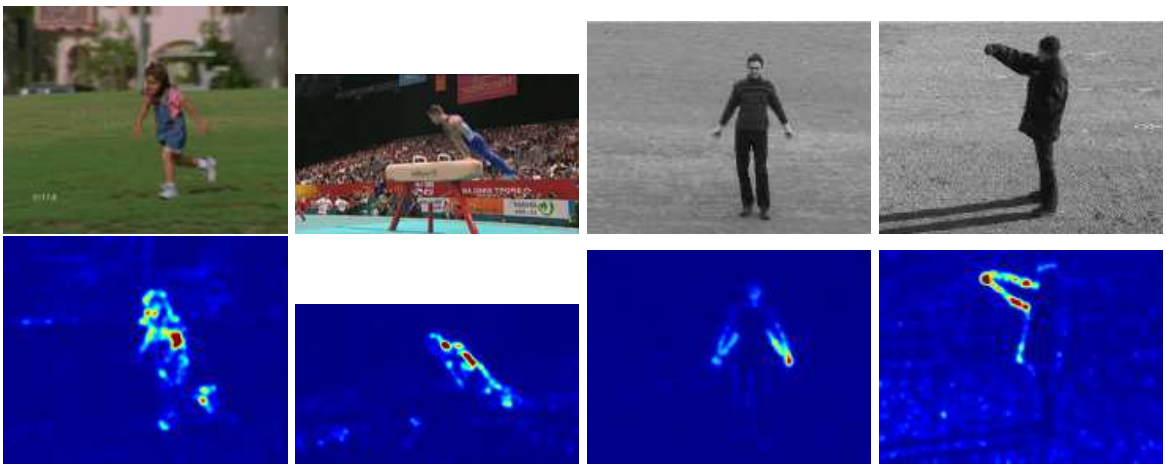


Figure 5.10: Some samples frames (left) from UCF sports action dataset (Column 1, 2) and KTH dataset (Column 3, 4) with their saliency maps (right).

5.5 Conclusion and Discussion

In this chapter, we proposed a novel approach for detecting spatiotemporal saliency, which was simple to implement and computationally efficient. The proposed approach was inspired by recent development of spectrum analysis based visual saliency approaches, where phase information was used for constructing the saliency map of the image. Recognizing that the computed saliency map captured the region of human's attention for dynamic scenes, we proposed two algorithms utilizing this saliency map for two important vision tasks. These approaches were evaluated on several well-known datasets with comparisons to the state-of-arts in the literature, where good results were demonstrated. For the future work, we will focus on theoretical analysis of the proposed method and the analysis on the selection of the window function.

Chapter 6

RELATIVE HIDDEN MARKOV MODELS FOR EVALUATING MOTION SKILLS

In our fourth problem, we work on how to learn a proper temporal model, which is essential to analysis tasks involving sequential data. In computer-assisted surgical training, which is the focus of this study, obtaining accurate temporal models is a key step towards automated skill-rating. Conventional learning approaches can have only limited success in this domain due to the lack of sufficient amount of data with accurate labels. In this chapter, we propose a novel formulation termed relative Hidden Markov Model and develop an algorithm for obtaining a solution under this formulation. The method requires only a sparse set of relative ranking information between input pairs (e.g., about 1000 out of 90000 pairs or about 1.1%), which are readily available from training sessions in the target application, hence alleviating the requirement on data labeling. The proposed algorithm learns a model from the training data so that the attribute under consideration is linked to the likelihood of the input, hence supporting comparing new sequences. For evaluation, synthetic data are first used to assess the performance of the approach, and then we experiment with real video data from a widely-adopted surgical training platform. Experimental results suggest that the proposed approach provides a promising solution to motion skill evaluation from video. To illustrate the generality of the method, we also report experiments on the task of emotion recognition from speech data.

6.1 Introduction

Human capability in mastering body motion is the key in domains such as sports, rehabilitation, surgery and dance. Computer-based approaches have been developed over the years for facilitating acquiring (e.g., training in sports and surgery) or regain-

ing (e.g., in rehabilitation) such motion-related skills by human subjects. One central task faced by systems using such approaches is the analysis of motion skills based on some temporal sensory data. With such analysis, skill metrics may be extracted and assigned to a given movement and feedback may accordingly be provided to the subjects for taking actions to improve the underlying skill. For example, Duan *et al.* (2008) utilized control trajectories and motion capture data for human skill analysis, Watanabe and Hokari (2006) reported motion skill analysis in sports using data from motion sensors, Suzuki *et al.* (2004) studied computational skill rating in manipulating robots, and Satoshi and Fumio (2010) considered hand movement analysis for skill evaluation in console operation.

Among others, surgery-related applications have attracted increasing interests, where motion expertise is the primary concern. To improve their motion expertise, surgeons often have to go through lengthy training processes. In recent years, simulation-based surgical training platforms have been developed and widely applied in surgical education. One prominent example is the Fundamentals of Laparoscopic Surgery (FLS) Trainer Box (`www.flsprogram.org`). With such platforms, it is possible to develop computational approaches to provide objective and quantifiable performance metrics, overcoming the shortcomings in traditional training that relies on costly practice of direct supervision by senior surgeons. Recognizing the sequential nature of motion data, many analysis approaches utilize state-transition models, such as the Hidden Markov Model (HMM). For example, Rosen *et al.* (2002) provided an HMM-based method to evaluate surgical residents' learning curve. The method first constructs different HMMs for each different levels of expertise, and then calculates a probability distance between the expert and a novice resident. The magnitude of the probability distance is used to rate the level of the novice resident. HMM was also adopted in Kahol *et al.* (2006) to measure motion skills in surgical tasks, where a recorded video is first segmented into basic gestures based on velocity and angle of movement, with segments of the gestures corresponding to the states of an HMM.

In Zhang and Li (2011), Hierarchical Dirichlet process hidden Markov model (HD-PHMM Fox (2009)) was utilized, which relaxed the requirement of predefining the number of the states for the model.

One practical difficulty in these approaches is that they require the skill labels for the training data since the HMMs are typically learned from sets of data streams with corresponding skill levels. Labeling the skill of a trainee is currently done by senior surgeons, which is not only a costly practice but also one that is subjective and less quantifiable. Thus it is difficult, if not impossible, to obtain a large amount of data with sufficiently reliable skill labels for HMM training. This problem has also been encountered in other fields such as image classification. For example, in Parikh and Grauman (2011), it was argued that using binary labels to describe images is not only too restrictive but also unnatural and thus relative visual attributes were used and classifiers were trained based on such features. Relative information has also been used in other applications, e.g., distance metric learning Schultz and Joachims (2004), face verification Kumar *et al.* (2009a), and human-machine interaction Parikh *et al.* (2012).

In this chapter, we propose a novel formulation termed *Relative Hidden Markov Model* and develop an algorithm for obtaining a solution under this model. The proposed method utilizes only a sparse set of relative ranking (based on certain attribute of interest, or motion skill in the surgical training application) between pairs of inputs, which is easier to obtain and often more consistent. This is especially useful for the applications like video-based surgical training, where the trainees go through a series of training sessions with their skills get improved over time, and thus the time of the sessions would already provide natural relative ranking of the skills at the corresponding time. The proposed algorithm effectively learns a model from the training data so that the attribute under consideration (i.e., the motion skill in our application) is linked to the likelihood of the inputs under the learned model. The learned model can then be used to compare new data pairs. For evaluation, we first

design synthetic experiments to systematically evaluate the model and the algorithm, and then experiment with real data captured on a commonly-used surgical training platform. The experimental results suggest that the proposed approach provides a promising solution to the real-world problem of motion skill evaluation from video.

The key contribution of the work lies in the novel formulation of learning temporal models using only a sparse set of relative information and the proposed algorithm for obtaining solutions under the formulation. A discussion of its relationship to the latent support vector machine is also provided to assist the understanding of why the proposed formulation is suitable for the proposed scenarios. Additional contributions include the specific application of the proposed method to the problem of video-based motion skill evaluation in surgical training, which has seen increasing importance in recent years. An earlier exposition of the proposed method can be found in Zhang and Li (2013). This current paper represents a full exploration of the method, including a new learning algorithm that is more efficient, new comparative analysis of the method, and new and updated experiments. Source code accompanying this paper is made publicly available to interested researchers for further exploration and comparison [1].

In the remainder of this chapter, we first review some of the related work in Sec. 6.2 and describe basic notations of the HMM in Sec. 6.3. The proposed method is then presented in Sec. 6.4, including a new algorithm for getting the solutions in Sec. 6.4.3 and a discussion of its relationship to latent support vector machine in Sec. 6.4.4. The proposed method is evaluated on three types of data in Sec. 6.5 including synthetic data (Sec. 6.5.1) and videos from surgical simulation systems (Sec. 6.5.2). We also present the experiment on the emotion recognition on speech data set to show the generality of the proposed method in domains other than surgical training. The paper is concluded in Sec. 6.7.

[1] The code is available at `www.public.asu.edu/~bli24/CodeSoftwareDatasets.html`

6.2 Related Work

In this section, we review two categories of existing work, discriminative learning for hidden Markov models and learning based on relative information, which are most related to our effort. Distinction between our proposed method and the reviewed work will be briefly stated.

Discriminative learning for HMM: Maximum-likelihood methods for learning HMM (e.g., the forward-backward algorithm) in general do not guarantee the discrimination ability of the learned models. To this end, several discriminative learning methods for HMM have been proposed. In Collins (2002), a discriminative training method for HMM was proposed based on perceptron algorithms. The methods iterates between the Viterbi algorithm and the additive update of the models. Hidden Markov Support Vector Machine (HM-SVM) was proposed in Altun *et al.* (2003), which combines SVM with HMM to improve the discrimination power of the learned model. These methods are "supervised" in nature, and thus the labeling of the state sequence is required for the training data, which limits their practical use. In Sloin and Burshtein (2008), another discriminative learning method for HMM was proposed, which only requires the labels of the training sequences. The method initializes the HMMs with maximum-likelihood method and then updates the models with SVM. One drawback is that, the updated models do not always lead to valid HMMs, which could be problematic for a physics-driven problem where the model states have real meanings (like the gesture elements in Kahol *et al.* (2006)). Our proposed method requires neither the labeling of the states nor the class label for the training sequences, which are difficult to obtain or even not accessible in applications. Instead, only a sparse set of relative ranking of the training data is used (e.g., about 1000 out of 90000 pairs or about 1.1%), and the resultant model is a valid HMM.

Learning with relative information: Several methods for learning with relative information have been proposed recently. In Schultz and Joachims (2004), a distance metric is learned from relative comparisons. Considering the limited train-

ing examples for object recognition, Wang *et al.* (2010) proposes an approach based on comparative objective similarities, where the learned model scores high for objects of similar categories and low for objects of dissimilar categories. In Kumar *et al.* (2009a), comparative facial attributes were learned for face verification. The method of Parikh and Grauman (2011) learns relative attributes for image classification and the problem is formulated as a variation of SVM. Similar idea was also been used in Parikh *et al.* (2012) for the purpose of human-machine interaction. In Kovashka *et al.* (2012), relative attributes feedback, e.g., "Shoe images like these, but sportier", is used to improve the performance of image search. Relative information between scene categories has also been used to enhance the performances of scene categorization in Kadar and Ben-Shahar (2012). These approaches are mostly for image-based attributes, whereas our current task is on modeling sequential data, for which it is natural to assume that the most relevant attributes (e.g., motion skills) are embedded in a temporal structure. This is what our proposed method attempts to address.

6.3 Basic Notations of HMM

In this section, we briefly describe HMM and introduce some basic notations that will be used later. An HMM can be defined by a set of parameters: the initial transition probabilities $\pi \in \mathbb{R}^{K\times 1}$, the state transition probabilities $A \in \mathbb{R}^{K\times K}$ and the observation model $\{\phi_k\}_{k=1}^{K}$, where K is the number of states. There are two central problems in HMM: 1) learning a model from the given training data; and 2) evaluating the probability of a sequence under a given model, i.e., the decoding problem.

In the **learning problem**, one learns the model (θ) by maximizing the likelihood of the training data ($\mathbb{X}$):

$$\theta^* : \max_{\theta} \prod_{\mathbf{X}^i \in \mathbb{X}} p(\mathbf{X}^i|\theta) \sim \max_{\theta} \sum_{\mathbf{X}^i \in \mathbb{X}} \log p(\mathbf{X}^i|\theta) \tag{6.1}$$

where $\mathbb{X}$ is the set of i.i.d. training sequences.

One efficient solution to the above problem is the well-known Baum-Welch algorithm Baum *et al.* (1970). Another scheme, namely the segmental K-means algorithm Juang and Rabiner (1990), may also be used to seek a solution, and it has been shown that the likelihoods under models estimated by either of the two algorithms are very close Juang and Rabiner (1990). When the training data include sequences of multiple categories, multiple models would be learned and each model will be learned from data of each category independently.

In the **decoding problem**, given a hidden Markov model, one needs to determine the probability of a given sequence $\mathbf{X}$ being generated by the model. Generally we are more interested in the probability associated with the optimal state sequence $(\mathbf{z}^*)$, i.e., $p(\mathbf{X}, \mathbf{z}^*|\theta) = \max_{\mathbf{z}} p(\mathbf{X}, \mathbf{z}|\theta)$. The optimal state path can be found via the Viterbi algorithm. To use HMM in classification, we first compute the probability of the given sequence drawn from each model, then we choose the model yielding the maximal probability.

6.4 Proposed Method

Based on the previous discussion, we are concerned with a new problem of learning temporal models using only a sparse set of relative information. This is a problem arising naturally in many applications involving motion or video data. In the case of video-based surgical training, the focus is on learning to rate/compare the performance of the trainees from recorded videos capturing their motion. To this end, in recognition of some fruitful trials of HMMs in this application domain, we propose to formulate the task as one of learning a *Relative Hidden Markov Model*, which not only maximizes the likelihood of the training data, but also maintains the given set of relative rankings of the input pairs. In its most basic form, the proposed model

can be formally expressed as (following the notations defined in Eqn. (1))

$$\begin{aligned} \theta \quad &: \quad \max_{\theta} \prod_{\mathbf{X}^i \in \mathbb{X}} p(\mathbf{X}^i|\theta) \\ \text{s.t.} \quad &\quad F(\mathbf{X}^i, \theta) > F(\mathbf{X}^j, \theta), \forall (i,j) \in \mathbb{E} \end{aligned} \tag{6.2}$$

where $F(\mathbf{X}, \theta)$ is a score function for data $\mathbf{X}$ given by model θ, which is introduced to maintain the relative ranking of the pair $\mathbf{X}^i$ and $\mathbf{X}^j$, $\mathbb{E}$ is the set of given pairs with prior ranking constraint and $\mathbb{F}$ is the set of given pairs required to have same response for the score function. $\mathbb{E}$ and $\mathbb{F}$ could be very sparse set compared with all pairs of training data. Different score functions may be defined, as described in the following subsections.

From this formulation, the difference between the proposed method and any of the existing HMM-based methods is obvious. In an existing HMM-based method, a set of models is trained using the training data of each category independently. That is, explicit class labels are required for each training sequence. The proposed model has the following unique features:

- The model does not require explicit class labels. What needed is only a relative ranking.
- The model explicitly considers the ranking constraint between given data pairs, whereas independently-trained HMMs in existing methods can't guarantee it.
- Only one model is learned for the entire set of data. There are two benefits: more data for training and less computation during testing.

Our method is also different from the existing work on learning with relative attributes in that it models sequential data and the relative ranking information is captured in a temporal dynamic model of HMM (albeit new algorithms are thus called for), which has demonstrated performance in modeling physical phenomena like human movements.

In the following subsections, we present two instantiations of the general model expressed in Eqn. (2), and develop the corresponding algorithms in each case. It will become clear that the first model (Sec. 6.4.1), while being intuitive, has some practical difficulties, which motivated us to develop the improved model of Sec. 6.4.2. Both models/algorithms are presented (and evaluated later in Sec.) for the progressive nature of the methods and for facilitating the understanding of the improved model and algorithm of Sec. 6.4.2, which is the recommended solution.

6.4.1 The Baseline Model

One intuitive choice of the score function in Eqn. (2) is the data likelihood, i.e., $F(\mathbf{X}^i, \theta) = p(\mathbf{X}^i|\theta)$. With this, the formulation in Eqn. (2) can be rewritten as

$$\begin{aligned} \theta \quad &: \quad \max_{\theta} \prod_{\mathbf{X}^i \in \mathbb{X}} p(\mathbf{X}^i|\theta) \\ \text{s.t.} \quad & \quad p(\mathbf{X}^i|\theta) > p(\mathbf{X}^j|\theta), \forall (i,j) \in \mathbb{E} \end{aligned} \tag{6.3}$$

It has been proved in Merhav and Ephraim (1991) that, the marginal likelihood is dominated by the likelihood with the optimal path and their difference decreases exponentially with regarding to the length (number of frames) of sequence. This idea was used in segmental K-means algorithm and similarly we can approximate the marginal data likelihood $p(\mathbf{X}|\theta)$ by the likelihood with optimal path $p(\mathbf{X}, \mathbf{z}^*|\theta)$ (when there is no ambiguity, we will use $\mathbf{z}$ for $\mathbf{z}^*$), which can be written as:

$$\log p(\mathbf{X}, \mathbf{z}|\theta) = \log p(\mathbf{X}_1|\phi_{\mathbf{z}_1}) + \log \pi(\mathbf{z}_1) + \sum_{t=2}^{T} [\log p(\mathbf{X}_t|\phi_{\mathbf{z}_t}) + \log \mathbf{A}(\mathbf{z}_t|\mathbf{z}_{t-1})] \tag{6.4}$$

For some observation models, e.g., multinomial (more details in Appendix C), we can write $\log p(\mathbf{X}^i, \mathbf{z}^i|\theta) = \theta^T h(\mathbf{X}^i, \mathbf{z}^i)$. Accordingly, Eqn. 6.3 can be finally written

as

$$\begin{aligned} \theta \quad &: \quad \max_{\theta \in \Omega} \theta^T \sum_{i:\mathbf{X}^i \in \mathbb{X}} h(\mathbf{X}^i, \mathbf{z}^i) \\ \text{s.t.} \quad & \quad \theta^T h(\mathbf{X}^i, \mathbf{z}^i) \geq \theta^T h(\mathbf{X}^j, \mathbf{z}^j) + \rho, \forall (i,j) \in \mathbb{E} \end{aligned} \tag{6.5}$$

where $\rho \geq 0$ defines the required margin between the logarithms of likelihood for a pair of data and Ω defines the set of valid parameters for the hidden Markov model, i.e.:

$$\begin{aligned} \theta(i) \leq 0 \quad ; \quad \sum_{i:\theta(i) \in \log(\pi)} e^{\theta(i)} = 1 \\ \sum_{i:\theta(i) \in \log(\mathbf{A}_j)} e^{\theta(i)} = 1 \quad ; \quad \sum_{i:\theta(i) \in \log(\phi_j)} e^{\theta(i)} = 1 \end{aligned} \tag{6.6}$$

where $i : \theta(i) \in \log(A_j)$ is the set of the indexes which corresponds to the j_{th} row of matrix A.

For the model in Eqn. 6.3, we assumed that every pair-wise ranking constraint provided in the sparse set is correct (or valid). However, in real data, there may be outliers in such training pairs. To handle this, we further introduce some slack variables ϵ and η, and accordingly Eqn. 6.5 can be written as following:

$$\begin{aligned} \theta \quad &: \quad \max_{\theta \in \Omega} \theta^T \sum_{\mathbf{X}^i \in \mathbb{X}} h(\mathbf{X}^i, \mathbf{z}^i) - \gamma \sum_{(i,j) \in \mathbb{E}} \epsilon_{ij} \\ \text{s.t.} \quad & \quad \theta^T [h(\mathbf{X}^i, \mathbf{z}^i) - h(\mathbf{X}^j, \mathbf{z}^j)] + \epsilon_{ij} \geq \rho, \forall (i,j) \in \mathbb{E} \\ & \quad \epsilon_{ij} \geq 0 \end{aligned} \tag{6.7}$$

where γ is the weight for the penalty term $\sum_{(i,j) \in \mathbb{E}} \epsilon_{ij}$. For initialization, we can set $\epsilon_{ij} = 0$. We will defer the optimization algorithm for Eqn. 6.7 to Sec. 6.4.3. After the model is learned, it can be used to a testing pair: For each sequence we evaluate the data likelihood via the Viterbi algorithm and use the logarithm of the data likelihood as the score of the data. By definition, the obtained scores can be used to compare the pair.

6.4.2 The Improved Model

In the model described in Eqn. 6.7, we compare the logarithm of the data likelihood, which is, according to Eqn. 6.4, roughly proportional to the length of the data. Thus a shorter sequence is likely to have a larger score. This means that the learned model would be biased towards the shorter sequences. If the observation describes a long, periodic event, e.g., repeating an action multiple times within a sequence, we may consider normalizing the logarithm of the data likelihood by the number of frames of the observation. However, this cannot be applied directly for non-periodic observations.

To overcome the above practical problem, we consider an improved version. Recall that in HMM, we classify a sequence based on the model with which the sequence gets the maximal likelihood, i.e., it is the ratio of data likelihood with different models that decides the label of the data. For example, if $\log \frac{p(\mathbf{X},\hat{\mathbf{z}}|\theta_1)}{p(\mathbf{X},\tilde{\mathbf{z}}|\theta_2)} > 0$, then we assign $\mathbf{X}$ to Model θ_1. Thus we propose to use the ratio of the data likelihoods of two HMMs as the score function, i.e., $F(\mathbf{X},\theta) = \log \frac{p(\mathbf{X},\hat{\mathbf{z}}|\theta_1)}{p(\mathbf{X},\tilde{\mathbf{z}}|\theta_2)}$, where we "partition" the original model into two models (or, effectively, we train a pair of HMMs simultaneously). This results in the following improved model:

$$\begin{aligned} \theta_1,\theta_2 \quad : \quad & \max_{\theta_1,\theta_2} \sum_{i\in\Xi_1} \log p(\mathbf{X}^i,\hat{\mathbf{z}}^i|\theta_1) + \sum_{j\in\Xi_2} \log p(\mathbf{X}^j,\tilde{\mathbf{z}}^j|\theta_2) - \gamma \sum_{(i,j)\in\mathbb{E}} \epsilon_{ij} \\ \text{s.t.} \quad & \log \frac{p(\mathbf{X}^i,\hat{\mathbf{z}}^i|\theta_1)}{p(\mathbf{X}^j,\tilde{\mathbf{z}}^j|\theta_2)} - \log \frac{p(\mathbf{X}^j,\hat{\mathbf{z}}^j|\theta_1)}{p(\mathbf{X}^j,\tilde{\mathbf{z}}^j|\theta_2)} + \epsilon_{ij} \geq \rho \\ & \epsilon_{ij} \geq 0 \end{aligned}$$

where Ξ_1 is the set of data associated with Model θ_1 (Ξ_2 for Model θ_2), $\hat{\mathbf{z}}^i$ is the optimal path for sequence x^i with Model θ_1 and $\tilde{\mathbf{z}}^i$ for optimal path with Model θ_2.

With $\log \frac{p(\mathbf{X}^i, \hat{\mathbf{z}}^i|\theta_1)}{p(\mathbf{X}^j, \tilde{\mathbf{z}}^j|\theta_2)} = \theta_1^T h(\mathbf{X}^i, \hat{\mathbf{z}}^i) - \theta_2^T h(\mathbf{X}^i, \tilde{\mathbf{z}}^i)$, we can rewrite the model in Eqn. 6.8 as the similar form in Eqn. 6.7:

$$\begin{aligned} \theta \quad : \quad & \max_{\theta \in \Omega} \theta^T \begin{bmatrix} \sum_{i \in \Xi_1} h(\mathbf{X}^i, \hat{\mathbf{z}}^i) \\ \sum_{j \in \Xi_2} h(\mathbf{X}^j, \tilde{\mathbf{z}}^j) \end{bmatrix} - \gamma \sum_{(i,j) \in \mathbb{E}} \epsilon_{ij} \\ \text{s.t.} \quad & \theta^T \begin{bmatrix} h(\mathbf{X}^i, \hat{\mathbf{z}}^i) - h(\mathbf{X}^j, \hat{\mathbf{z}}^j) \\ h(\mathbf{X}^j, \tilde{\mathbf{z}}^j) - h(\mathbf{X}^i, \tilde{\mathbf{z}}^i) \end{bmatrix} + \epsilon_{ij} \geq \rho \\ & \epsilon_{ij} \geq 0 \end{aligned} \tag{6.8}$$

where $\theta = [\theta_1^T, \theta_2^T]^T$. The optimization algorithm for Eqn. 6.8 will be presented in Sec. 6.4.3. After we learn the model with the improved algorithm, we can apply it to a given pair by first computing their likelihoods with respect to the "sub-models" given by θ_1 and θ_2 (with the Viterbi algorithm), and then we use the logarithm of the ratio of the data likelihoods as the score to rank/compare the pair.

The learned models θ_1 and θ_2 serve as a unified model to rank the data. We may view them as the centers of two clusters, where the distances of the data to those two centers can be related to the ranking score.

It needs to be emphasized that the improved model is not equivalent to a supervised HMM with two classes. In a 2-class HMM setting, two models will be independently trained with their respective training sets. Here, the proposed model trains two "sub-models" jointly with only relative ranking constraints. Specifically, if there is no further information for Ξ, we could assume that $\Xi_1 = \{i|(i,j) \in \mathbb{E}, \forall j\}$ and $\Xi_2 = \{j|(i,j) \in \mathbb{E}, \forall i\}$, and thus there could be overlaps between Ξ_1 and Ξ_2 (which will become clear in the experiment with synthetic data in Sec. 6.5). This situation not even allowed by a supervised HMM setting. We don't require any extra properties for Ξ_1 and Ξ_2, e.g., balances.

6.4.3 Algorithms for Updating the Model

One important step of both the baseline algorithm and the improved algorithm is updating the models, as formulated in Eqn. 6.7 and Eqn. 6.8 accordingly. It

is a nonlinear programming problem (due to the nonlinear equality constraint). In our previous paper, we solve it by the primal-dual interior point method, which is of dimension $K(1+K+D)+|\mathbb{E}|$ (or $2K(1+K+D)+|\mathbb{E}|$) with $2|\mathbb{E}|+K(1+K+D)$ (or $2|\mathbb{E}|+2K(1+K+D)$) linear inequality constraints and $1+K+D$ (or $2(1+K+D)$) nonlinear equality constraints for the baseline model (or the improved model). Although the Hessian matrix is diagonal, the computational cost could be still very high when there are a large number of training pairs. In this section, we propose to use a new algorithm by utilization the special structure of the problems in Eqn. 6.7 and Eqn. 6.8.

Eqn. 6.7 (similarly for Eqn. 6.8) can be written in the following form:

$$\begin{aligned} \theta, \epsilon \quad &: \quad \min_{\theta,\epsilon} \mathbf{f}^T\theta + \gamma 1^T\epsilon \\ \text{s.t.} \quad &: \quad \mathbf{A}\theta + \epsilon \leq \rho \\ &\quad\quad \mathbf{C}e^{\theta} = 1 \\ &\quad\quad \theta \leq 0; \epsilon \geq 0 \end{aligned} \tag{6.9}$$

For example, for Eqn. 6.7, we have $\mathbf{f} = -\sum_{\mathbf{X}^i \in \mathbb{X}} h(\mathbf{X}^i, \mathbf{z}^i)$, $\mathbf{A}$ and $\mathbf{C}$ are constructed according to Eqn. 6.7 and 6.6.

Eqn. 6.9 is a nonlinear programming problem (due to the nonlinear equality constraint). To solve this problem, we first introduce a slack variables ϕ, where $\log\phi = \theta$. Then Eqn. 6.9 can be rewritten into the following problem:

$$\begin{aligned} \theta, \epsilon, \phi \quad &: \quad \min_{\theta,\epsilon,\phi} \mathbf{f}^T\theta + \gamma 1^T\epsilon \\ \text{s.t.} \quad &: \quad \mathbf{A}\theta + \epsilon \leq \rho \\ &\quad\quad \mathbf{C}\phi = 1 \\ &\quad\quad \log\phi = \theta \\ &\quad\quad \theta \leq 0; \epsilon \geq 0; 0 \leq \phi \leq 1 \end{aligned} \tag{6.10}$$

According to Eqn, 6.10, ϕ will be a valid hidden Markov model (or hidden Markov model pairs $[\phi_1, \phi_2]$ for improved model). We then apply the Augmented Lagrange

multiplier method to the equality constraint $\log \phi = \mathbf{u}$ of the problem in Eqn. 6.10:

$$\begin{aligned} \theta, \epsilon, \phi \quad &: \quad \min_{\theta,\epsilon,\phi} \mathbf{f}^T\theta + \gamma 1^T\epsilon + < \lambda, \theta - \log\phi > + \frac{\mu}{2}\|\theta - \log\phi\|_2^2 \\ \text{s.t.} \quad &: \quad \mathbf{A}\theta + \epsilon \leq \rho \\ &\quad \mathbf{C}\phi = 1 \\ &\quad \theta \leq 0; \epsilon \geq 0; 0 \leq \phi \leq 1 \end{aligned} \tag{6.11}$$

where λ is the Lagrange multiplier and μ is some non-negative constant. In Eqn. 6.11, the nonlinear equality constraint is removed.

Eqn. 6.11 can be solved via block coordinate descent by iterating between the following two sub-problems:

Sub-problem 1: fix ϕ to solve θ and ϵ, which is

$$\begin{aligned} \theta, \epsilon \quad &: \quad \min_{\theta,\epsilon,\phi} \mathbf{f}^T\theta + \gamma 1^T\epsilon + < \lambda, \theta - \log\phi > + \frac{\mu}{2}\|\theta - \log\phi\|_2^2 \\ \text{s.t.} \quad &: \quad \mathbf{A}\theta + \epsilon \leq \rho \\ &\quad \theta \leq 0; \epsilon \geq 0 \end{aligned} \tag{6.12}$$

It is a quadratic programming problem with linear inequality constraints.

Sub-problem 2: fix θ and ϵ to solve ϕ, which is

$$\begin{aligned} \phi \quad &: \quad \min_{\phi} < \lambda, \theta - \log\phi > + \frac{\mu}{2}\|\theta - \log\phi\|_2^2 \\ s.t. \quad &\quad \mathbf{C}\phi = 1 \\ &\quad 0 \leq \phi \leq 1 \end{aligned} \tag{6.13}$$

It is a nonlinear problem with linear constraints.

Given the special structures of $\mathbf{C}$, where each column has one and only one element being nonzero (recall Eqn. 6.6), Sub-problem 2 can be separated into a set of smaller problems:

$$\begin{aligned} \phi^k \quad &: \quad \min_{\phi^k} < \lambda^k, \theta^k - \log\phi^k > + \frac{\mu}{2}\|\theta^k - \log\phi^k\|_2^2 \\ s.t. \quad &\quad \mathbf{1}^T\phi^k = 1 \\ &\quad 0 \leq \phi^k \leq 1 \end{aligned} \tag{6.14}$$

where k is the set of indexes of columns, whose values are nonzero at k_{th} row of $\mathbf{C}$. Those smaller problems are again a nonlinear problem with linear constraint, whose dimensions are only K (number of states) or D (number of feature dimension).

To solve this problem we can use the primal-dual interior point method, whose gradient and hessian are computed as

$$\begin{aligned} J &= \frac{-\lambda^k + \mu^k \log \phi^k - \mu^k \theta^k}{\phi^k} \\ H &= \Lambda(\frac{\lambda^k - \mu \log \phi^k + \mu\theta^k + \mu}{\phi^k \cdot \phi^k}) \end{aligned}$$

where $\Lambda(\cdots)$ converts a vector to a diagonal matrix. In addition, we can compute the starting point of the problem in Eqn. 6.14 as: by taking the gradient of the objective function with regard to $\log \phi^k$, we have $-\lambda^k + \mu(\log \phi^k - \theta^k) = 0$, i.e., $\phi^k = e^{(\theta^k + \frac{\lambda^k}{\mu})}$. The linear constraint can be solved by simply projection, i.e., $\phi^k = \frac{1}{N} e^{(\theta^k + \frac{\lambda^k}{\mu})}$, where $N = \sum e^{(\theta^k + \frac{\lambda^k}{\mu})}$.

Finally, we briefly summarize the algorithms for the baseline model (Eqn. 6.7) and the improved model (Eqn. 6.8) below (noting the similarity in form of the algorithms and thus putting them compactly together):

Algorithm for the Baseline (Improved) Model

Input: $\mathbb{X}$, $\mathbb{E}$, ρ, γ, σ (, Ξ_1 and Ξ_2)

Output: ϕ

Initialization: Initialize ϕ (or ϕ_1 and ϕ_2) via ordinary HMM learning algorithm, $\lambda = \frac{\log \theta}{|\theta|_2}$ and $\mu = \frac{1.25}{|\theta|_2}$;

while not converged **do**

Compute optimal path $\mathbf{z}$ (or $\hat{\mathbf{z}}$ and $\tilde{\mathbf{z}}$) for each sequence with ϕ (or ϕ_1 and ϕ_2);

solve Sub-problem 1;

solve Sub-problem 2;

update $\lambda = \lambda + \mu(\theta - \log \phi)$ and $\mu = \mu \times \sigma$;

check convergence;

end while

According to Bertsekas (1982), the proposed method will converge to the local minimum of the problem in Eqn. 6.9. And for convergence, we check $\frac{\|\theta - \log \phi\|_2}{\|\theta\|_2}$. If it is smaller than some value, e.g., 10^{-6}, the algorithm will be terminated. In initialization, $|\theta|_2$ is the vector L_2 norm of of θ.

Remarks on the Parameters: The parameter γ controls the weight of the penalty term with the slack variables, which is similar to the functionality of C in support vector machines Chang and Lin (2011). The parameter ρ controls the desired gap of the score of two data, i.e., $\frac{p(\mathbf{X}^i, \mathbf{z}^i|\theta)}{p(\mathbf{X}^j, \mathbf{z}^j|\theta)} \geq e^{\rho}\ \forall (i,j) \in \mathbb{E}$ in the baseline model and $\frac{p(\mathbf{X}^i, \hat{\mathbf{z}}^i|\theta_1)}{p(\mathbf{X}^i, \tilde{\mathbf{z}}^i|\theta_2)} \frac{p(\mathbf{X}^i, \tilde{\mathbf{z}}^j|\theta_2)}{p(\mathbf{X}^i, \hat{\mathbf{z}}^j|\theta)} \geq e^{\rho}\ \forall (i,j) \in \mathbb{E}$ in the improved model. In Sec. 6.5.1, we will evaluate different parameter settings (Fig. 6.2), which leads us to set $\gamma = 1000$ and $\rho = 10$ in our final experiments. The parameter σ controls convergence speed of the algorithm in Sec. 6.4.3, which should be a positive number and larger than 1. σ is typically within $1.1 - 1.5$, and 1.25 is used in this paper.

The proposed algorithm, compared with the one used in Zhang and Li (2013), has lower computational cost, due to the removal of the nonlinear equality constraint with augmented Lagrange multiplier. For Sub-problem 1, the quadratic term is a diagonal matrix and many solvers (e.g., CPLEX) can solve it quite efficiently. Sub-problem 2 is a nonlinear minimization problem with linear equality constraints; however, it can be decomposed into several smaller problems.

A comparison between the method in Zhang and Li (2013) and the proposed method for updating the baseline model is shown in Tab. 6.1. In Sec. 6.5.1, we will also compare the computational time of those two methods under varying $\mathbb{E}$ on synthetica data (Fig. 6.6).

	Method in Zhang and Li (2013)	Proposed Method	
		Sub-problem 1	Sub-problem 2
Problem Size	$K(1+K+D)+\|\mathbb{E}\|$	$K(1+K+D)+\|\mathbb{E}\|$	K(or D)
# Linear Const.	$2\|\mathbb{E}\|+K(1+K+D)$	$2\|\mathbb{E}\|+K(1+K+D)$	1+2K(or 1+2D)
# Nonlinear Const.	$1+K+D$	0	0

Table 6.1: Comparing the method in Zhang and Li (2013) and the proposed method for updating the baseline model, with regarding to the problem size, number of linear constraints and nonlinear constraints. For Sub-problem 2 of the proposed method, it can be divided into several smaller problems.

6.4.4 Relationship to Existing Methods

The proposed method is related to latent support vector machine Felzenszwalb *et al.* (2010). Given a training set of input-output pairs $\{(x_i, y_i)\}_{i=1}^n$, where $y_i \in \{-1, 1\}$, Latent SVM tries to learn a predictor of the form:

$$f_w(x) = \max_z w^T \Psi(x, z) \tag{6.15}$$

where w is the parameter of the predictor, $\Psi(x, z)$ is the feature mapping function and z is the latent variable. The training stage of Latent SVM can be formulated as the following problem:

$$\min_w \frac{1}{2}\|w\|_2^2 + C\sum_i \max\left(0, 1 - y_i f_w(x_i)\right) \tag{6.16}$$

Latent SVM is a non-convex problem, as the latent variable is unknown, and the coordinate descent approach is used for solving this problem.

Given a training set $\{(x_i, y_i)\}_{i=1}^n$, where $x_i = (\mathbf{x}_i^L, \mathbf{x}_i^R)$ is a pair of sequences and $y_i \in \{-1, 1\}$ is the ranking of the pair, by defining the feature mapping function as $\Psi(x_i, z_i) = [h(\mathbf{x}_i^L, \mathbf{z}_i^L) - h(\mathbf{x}_i^R, \mathbf{z}_i^R)]$ and latent variable $z_i = (\mathbf{z}_i^L, \mathbf{z}_i^R)$ is a pair of state sequences for the pair of sequences $x_i = (\mathbf{x}_i^L, \mathbf{x}_i^R)$ accordingly, we have

$$\begin{aligned} & \min_w \frac{1}{2}\|w\|_2^2 + C\sum_i \epsilon_i \\ \text{s.t.} \quad & y_i \max_{\mathbf{z}_i^L, \mathbf{z}_i^R} \left\{w^T[h(\mathbf{x}_i^L, \mathbf{z}_i^L) - h(\mathbf{x}_i^R, \mathbf{z}_i^R)]\right\} + \epsilon_i \geq 1 \\ & \epsilon_i \geq 0 \end{aligned} \tag{6.17}$$

We can find that Eqn. 6.17 is similar to our baseline model (Eqn. 6.7), except for the following differences.

1. In Eqn. 6.17, the L_2 norm is applied to the parameter of the predictor w (which is related to the margin). In the proposed methods we require w to be a valid hidden Markov model while defining a fixed-margin, i.e., ρ. Thus the proposed method can always guarantee the learned model is a valid hidden Markov model.

2. In Eqn. 6.17, the two state sequences z (i.e., the latent variables) are optimized jointly, where no known efficient solution is available. In the proposed method, the two state sequences are optimized separately with regarding to the likelihood, which can be solved efficiently via dynamic programming (i.e., the Viterbi algorithm);

3. Given the model learned by the latent SVM (Eqn. 6.7), we can only rank a pair of sequences. However, the model learned by the proposed method is capable of not only ranking a pair of sequences but also assigning a score for each sequence.

Those differences make the proposed method (both the baseline model and the improved model) more suitable for modeling the sequential data, e.g., video, speech.

6.5 Experiments

In this section, we evaluate the proposed methods, including the baseline method and the improved method, using both synthetic data (Sec. 6.5.1) and realistic data collected from the surgical training platform FLS box (Sec. 6.5.2). The performance of the proposed methods is compared with a supervised 2-class HMM. (Lacking a comparative approach in the literature that is both unsupervised and works with only relative rankings, this is believed to be a reasonable way of generating a reference point to assess the proposed methods.)

6.5.1 Evaluation with Synthetic Data

To evaluate the proposed method, we generate synthetic data: we first generate six different HMMs (θ_1 to θ_6, which are referred as data-generating models), from each of which we draw 200 sequences, with the length being uniformly distributed between 80 to 120. Each data-generating model has five states. For the sequences from each data-generating model, we randomly assign 50 of them to the training set and the remaining to the testing set. We assume there exists a score function such that $F(\mathbf{X}^i) > F(\mathbf{X}^j)$ if and only if $\mathbf{X}^i \sim \theta_k$, $\mathbf{X}^j \sim \theta_l$ and $k < l$. That is, the sequences from a data-generating model with a lower index are viewed to have a higher score (or ranking) than those from a data-generating model with a higher index. A set of pairs $\{(i,j)|\mathbf{X}^i \sim \theta_k, \mathbf{X}^j \sim \theta_{k+1}, k = 1, \cdots, 5\}$ are then formed accordingly, some sparse subset of which are then randomly selected as the training pairs $\mathbb{E}$.

We use the proposed methods and also HMM to learn models from the training pairs. For the HMM algorithm and the improved method, we initialize the two sets as $\Xi_1 = \{i|(i,j) \in \mathbb{E}, \forall j\}$ and $\Xi_2 = \{j|(i,j) \in \mathbb{E}, \forall i\}$. Note, the data generated from data-generating Models $\theta_2 \sim \theta_5$ could be included in both Ξ_1 and Ξ_2. Thus existing discriminative learning methods for HMM could not be applied here.

The learned models are then used to evaluate the testing set, i.e., how many testing pairs that they rank the same as ground truth. The result of the methods with different number of training pairs is summarized in Fig. 6.3, where due to the computational time it takes, we don't have the results for the baseline method when there are more than 3750 training pairs.. From Fig. 6.3, we can find that the improved method achieves the best results on both the training set and the testing set; and the HMM method gives the worse result. In addition, the performance of both of the proposed methods stabilized after certain number of training pairs. However the performance of the HMM method drops dramatically when the number of training pairs reaches about 6250. It can be explained by that the two HMMs share a lot of common data (for those generated by $\theta_2 \sim \theta_5$) and the models are trained

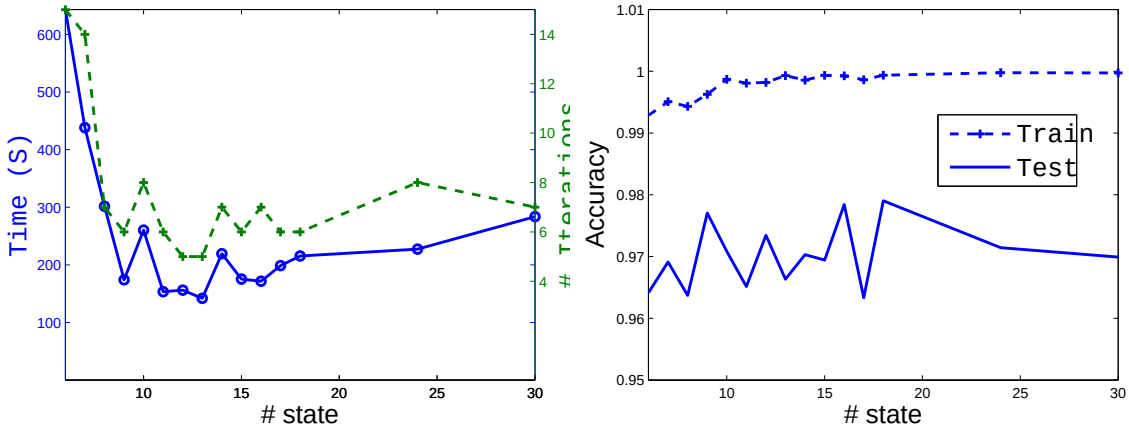
independently without consideration of their discrimination ability. Normalizing the logarithm of data likelihood does not improve the performance of baseline method, which could be explained by that, all the sequences have roughly the same length, i.e., $80 \sim 120$. Fig. 6.4 shows the logarithm of the data likelihood ratio with the models learned by the improved method, when about 1250 training pairs are provided. This clearly demonstrates that, although we formed the training pairs only with data from data-generating models of adjacent indexes (i.e., i and $i+1$), the learned model is able to recover the strict ranking of the original data.

Convergence and Speed For empirically understanding the convergence behavior of the improved method, we plot in Fig. 6.5 the objective value in the model as a function of the number of iterations. We can find that the improved method converges fairly quickly (within about 14 iterations) and the value of the objective function monotonically increases.

We also compare the computational time of the optimization method in Zhang and Li (2013) (shown as red curve) and the proposed optimization method (in Sec. 6.4.3 and shown as green curve) in solving the improved model under varying number of training pairs in Fig. 6.6. In Zhang and Li (2013), a primal-dual interior point method is utilized to update the model; while in this paper, we design an augmented Lagrange multiplier method which utilizes the special structure of the objective function of the problem. From the plot, we can find that the proposed optimization method has a much lower computational cost than the one proposed in Zhang and Li (2013).

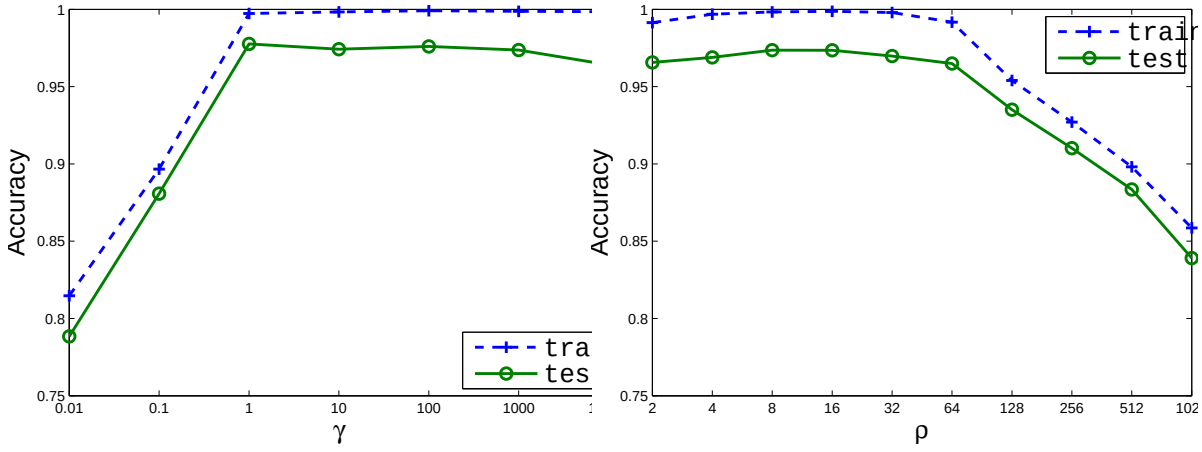
Parameter Selection: to understand the effect of parameters to the performances of the improved method, including accuracy and computation cost, we evaluate it with varying combination of parameters. First we learn the model with varying numbers of states (K), from 6 to 30. The result is shown Fig. 6.1. From Fig. 6.1(b), we can find that, though the accuracy for the training data increases with the number of states, the accuracy for testing doesn't following this trend, which indicates a potential risk of over-fitting. The computational time and number of iteration until

convergence get minimized when the number of states is $11 - 13$. We also do experiment with different combinations of γ (controlling the weight of the penalty term with slack variables) and ρ (controlling the margin of the model), where the experiment result is shown in Fig. 6.2. From this experiment we can find that, $\gamma \in [1, 1000]$ and $\rho \in [4, 32]$ are good choices.



a b

Figure 6.1: The experiment result with different numbers of states: (a) the computational time (blue solid curve) and number of iterations needed for convergence (green dashed curve); (b) the accuracy of the improved method. The X-axis is the number of states.



a b

Figure 6.2: The accuracy of the improved method: (a) with different γ (ρ is fixed to 10), which controls the weight of the penalty term with slack variables; (b) with different ρ (γ is fixed to 1000), which controls the margin of the learned models.

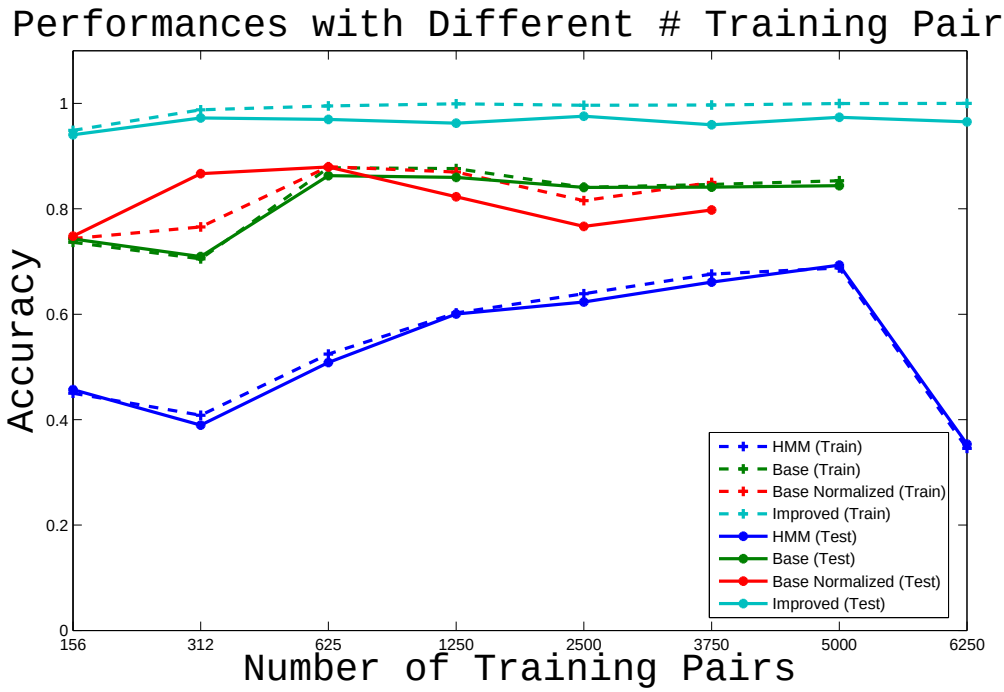


Figure 6.3: The results of four methods on training set (dashed curve) and testing set (solid curve) with different numbers of training pairs.

It is obvious from this experiment that the sequences are different from (or similar to) each other only because they are from different (or the same) data-generating models, whereas their relative ranking can be arbitrarily defined. In the end, the proposed methods will learn a temporal model to reflect the defined rankings. This suggests that, as long as we can assume there are some data-generating models for the given sequential data, we can use the proposed methods to learn a relative HMM. This is the basis for applying the approach to the surgical training data in the following sub-section, where it is reasonable to assume that movement patterns of subjects with different skill levels may be modeled by different underlying HMMs while the ranking can be based on the time of training, which reflects the skill level of the subject at the time.

6.5.2 Skill Evaluation Using Surgical Training Video

We now evaluate the proposed method using real videos captured from the FLS trainer box, which has been widely used in surgical training. The data set contains 546

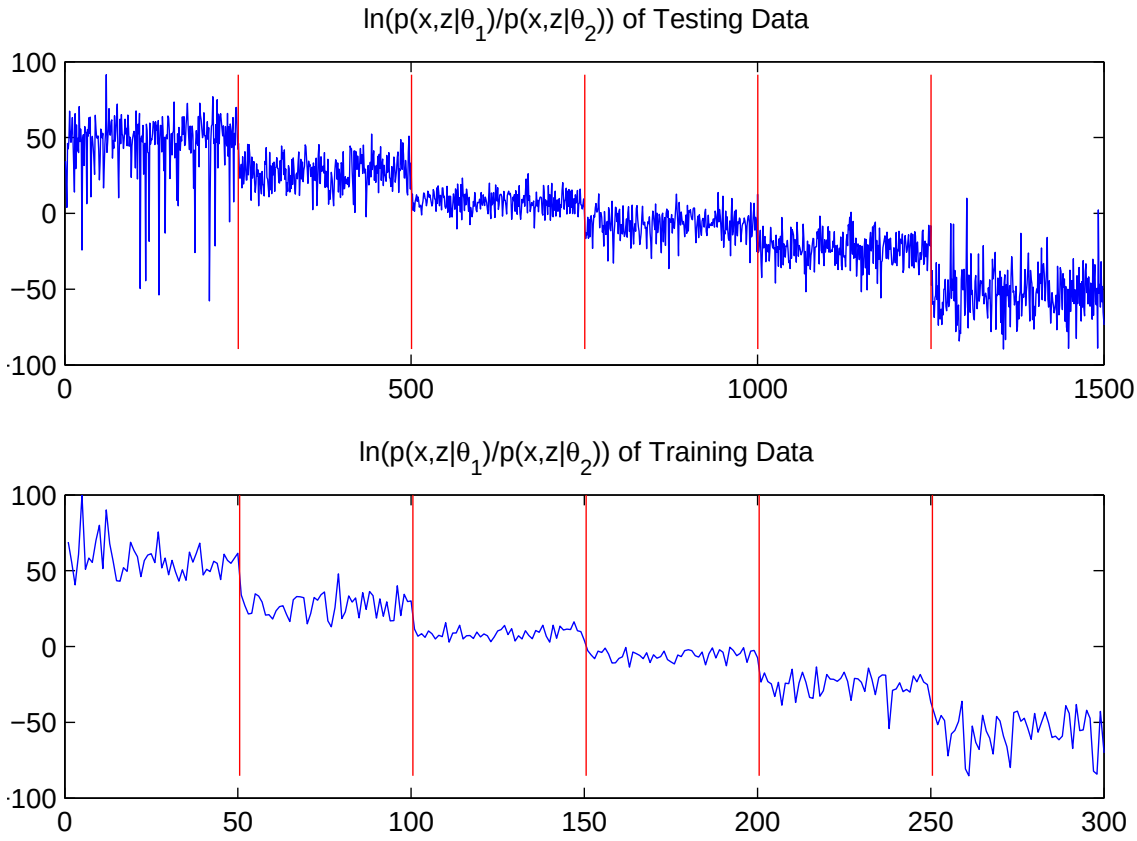


Figure 6.4: The logarithm of the data likelihood ratio with the models learned by the improved method. Top: the result for the testing set. Bottom: the result for the training set. The data are grouped (as the section partitioned by the red lines) according to the data generation model from which they are synthesized.

videos captured from 18 subjects performing the "peg transfer" operation, which is one of the standard training tasks a resident surgeon needs to perform and pass. The number of frames in each video varies from 1000 to 6000 (depending on the trainees' speed in completing a training session). The data set covers a training period of four weeks, with every trainee performing three sessions each week.

In the training, the subject needs to lift six objects (one by one) with a grasper by the non-dominant hand, transfer the object midair to the dominant hand, and then place the object on a peg on the other side of the board. Once all six objects are transferred, the process is reversed, and the objects are to be transferred back to the original side of the board. The videos capture the entire process inside the trainer box, showing how the tools and objects are moved by the subject. The motion skill is

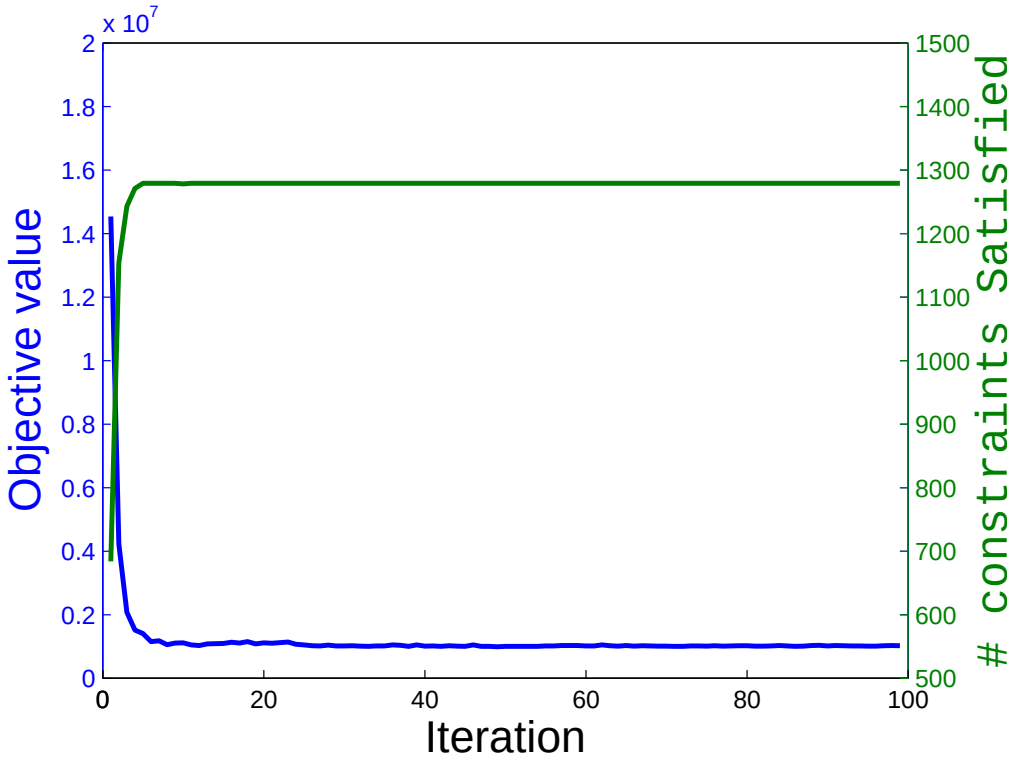


Figure 6.5: The convergence behavior of the improved method, where around 1250 training pairs were used. The blue curve/axis shows the value of the objective function, and the green curve/axis shows the number of constraints satisfied.

related to how well the subjects perform in such operation. In the existing practice, senior surgeons rate the performance of the trainees based on such videos. Our goal is to perform the rating automatically with the proposed model.

Based on the reasonable assumption that the trainees improve their skills over time (which is the whole point of having the resident surgeons going through the training before taking the exam), the time of recording is used to rank the recorded videos **within each subjects' corpus** (i.e., a later video is associated with a better skill). Other than this relative ranking, there are no other labels assumed for the video, e.g., there is no rank information between videos of different subjects (which would be hard to obtain anyway, since there are no clearly-defined skill levels for a group of trainees with diverse background). Based on this, our training set is built as we first build a set as $\{(i, j) : i$ is the video for the last week , j is the video of first week $\forall subject\}$,

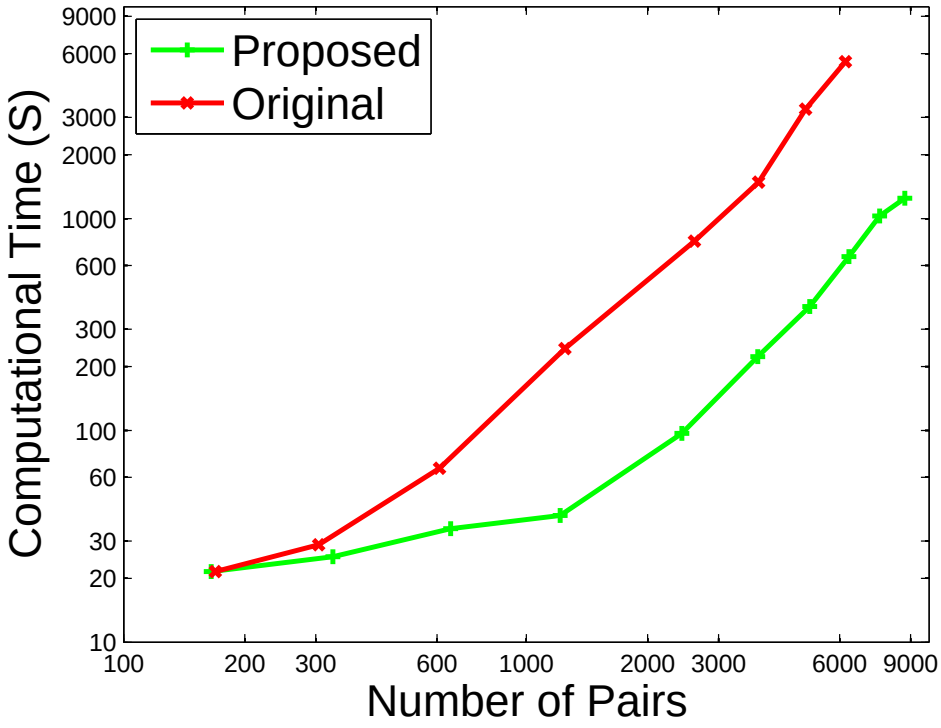


Figure 6.6: The computation time for solving the improved model with the method proposed in Zhang and Li (2013) (red curve) and the method proposed in Sec. 6.4.3 (green) under varying number of training pairs. For illustration purpose, we use log-log plot, where X-axis is the number of training pairs (from around 125 to around 9000) and Y-axis is the computation time in unit second (from about 20 to around 6000). The time is measured in Matlab on a dual-core PC platform.

then from it, we randomly pick 300 pairs as the training set. The training set is very sparse compared with over 8000 pairs.

Feature Extraction: we use the "bag of words" approach for feature extraction from the videos as follows. The spatiotemporal interest point detector Laptev (2005) is applied to obtain the histogram-of-gradient (HoG) features. K-means ($k = 100$) is then used to build a code book for the descriptors of the interest points. Finally, the code book is used to obtain a histogram of interest points for each frame, and thus each video is represented as a sequence of histograms. This representation, compared with the existing way of using bag of words in action recognition, i.e., transforming each video into a single histogram, can better capture the temporal information of the data. For all three methods, we set the number of states to ten.

After learning the models from the training data, we compute the score of the test data as the logarithm of data likelihood (for the baseline method) or the logarithm of the data likelihood ratio (for the improved method and the HMM). We compare these scores for each pair of the testing data (within each subject) and compute the percentage of correctly labeled pairs (recall that, we use their time of recording as ground truth). The result is summarized in Tab. 6.2, where the improved method obtained a significantly better result than the other approaches. Surprisingly, the baseline method even performed slightly worse than the HMM method. This is largely due to the wide range of variations of the length of the input sequences. Fig. 6.7 shows the computed scores with the learned models, where for better illustration purpose we group them by their subject ID and within each subjects' corpus we sort the videos by their recording time. From the figure, we can find that the improved method (bottom) reveals a more clear trend for the data than both the HMM method (top) and the baseline method (middle), i.e., the scores of the data increase over times (X-axis) for each subject (segments within the red lines). It is worth emphasizing that only one joint model is learned from ranked pairs of subjects with potentially varying skill levels. Still the learned model is able to recover the improving trend, independent of the underlying skill levels.

It is also interesting to look at what the jointly-learned models look like in the proposed approach. Fig. 6.8 depicts the two models learned by the improved method in this real-data based experiment. From the figure, we can see that the two models have different transition patterns. For example, the transition from State 8 to States 2 and 5 are only observed in Model 1. This may be linked to different motion patterns for data of different surgical skills.

6.6 Emotion Recognition from Speech Data

Although the proposed approach was evaluated above in the context of motion skill analysis in surgical training, the approach itself is general and applicable for other

Method	HMM	Baseline	Improved
# Pairs	6363	6215	**6993**
Accuracy	79.39%	77.54%	**87.25%**

Table 6.2: The result for experiment on evaluating surgical skills. There are 8015 pairs in total (only 300 for training), excluding the comparisons among data of different subjects.

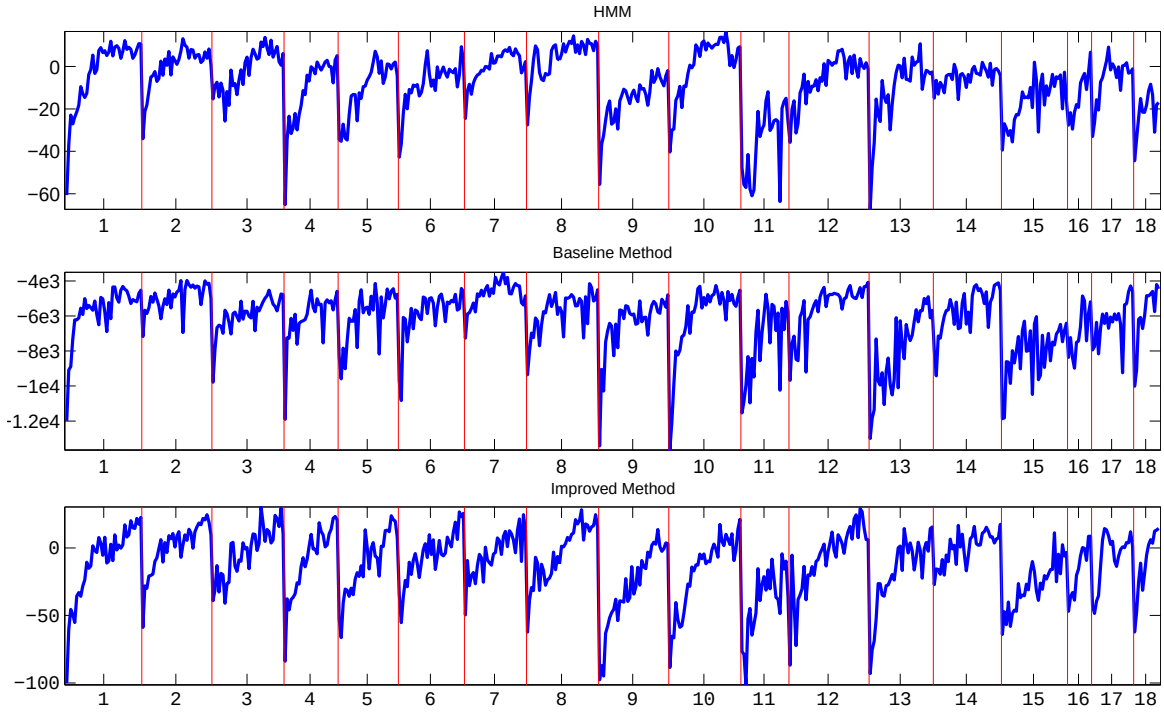


Figure 6.7: Top: the logarithm of the data likelihood ratio from two models learned by HMM. Middle: the logarithm of data likelihood with the model learned by the baseline method. Bottom: the logarithm of the data likelihood ratio with the models learned by the improved method. The red vertical lines separate the data of different subjects, where X-axis is the corresponding subject ID. Within each subjects' corpus, the videos are sorted according to their time of recording.

applications involving temporal data. To show that the proposed method can also be used to solve temporal inference problems other than motion skill assessment, we now consider the problem of speech-based emotion recognition. Recognizing the emotional state of the speakers has received quite some interests from researchers, due to its broad applications. For example, in human-machine interaction, better responses can be made if the emotional state of the human can be recognized. Existing work on this in the literature mainly focuses on developing models for assigning the labels like "pleasing", "angry" and "neural" to the data, e.g., Nwe *et al.* (2003), Schuller *et al.*

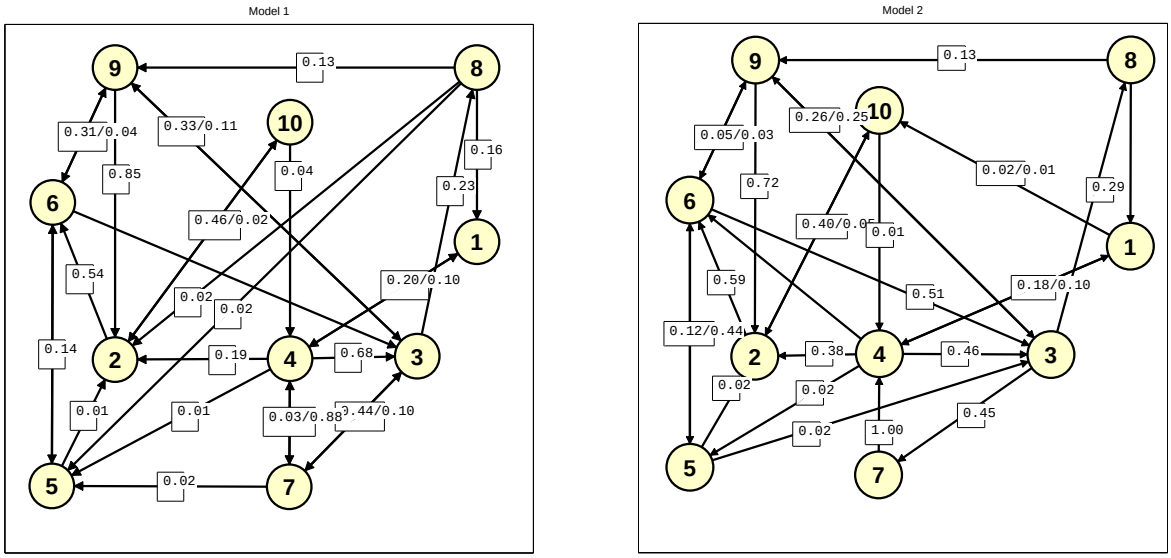


Figure 6.8: The two component models (Model 1 for Ξ_1 and Model 2 for Ξ_2) learned by the improved method, where we only draw the edges with a transition probability larger than 0.01 and ignore self-transitions. The number attached to each edge indicates the transition probability.

(2003), El Ayadi *et al.* (2011), Tarasov and Delany (2011). Most of those efforts are supervised in natural, i.e., the ground truth labeling for the training data is required. For example, Kim *et al.* (2004) used support vector machines, Nwe *et al.* (2003) used hidden Markov models, both utilizing fully-labeled data. The ground truth data typically require manual labeling by human, which is an error-prone process especially if absolute labels must be assigned to ambiguous data. With the proposed model, we can support learning with only relative labels like "Audio a is more pleasing than Audio b", which is easier to obtain and also less error-prone.

In this experiment, we use Utsunomiya University Spoken Dialogue Database For Paralinguistic Information Studies (UUDB)Mori *et al.* (2008), which contains 4840 assets labeled across six dimensions (pleasantness, arousal, dominance, credibility, interest and positivity) on a scale of 1 to 7. The ground truth is based on the average of scores of three annotators. For experiment, we pick the assets which are longer than 1 second to ensure the effectiveness of emotional recognition, which results in 991 assets, where half of the data are used for training and the remaining for testing. For generating the training set of pairs, we randomly picks 1000 pairs from the training

Dimension	Improved	Baseline	HMM
Pleasantness	**77.30**%	57.96%	75.05%
Arousal	**86.95**%	55.74%	69.55%
Dominance	**87.95**%	63.04%	77.32%
Credibility	**76.68**%	55.11%	71.74%
Interest	**81.90**%	62.56%	78.07%
Positivity	**74.99**%	67.84%	70.36%
Average	**81.28**%	53.14%	73.72%

Table 6.3: The result for experiment on UUDB datasets. We evaulate the accuracy of ranking pairs with the learned models compared with the ground truth ones.

assets, which is very sparse considering total of about $1000,000$ pairs. Note that, we say two assets are similar, if the difference of the labeled scores of two assets is within the range of $(-1, 1)$. The data can be downloaded at `http://uudb.speech-lab.org`.

For feature extraction, we use Hidden Markov Model Toolkit (HTK)Woodland *et al.* (1994), where the MFCC coefficients are extracted with the following configurations: sampling rate is 100 HZ, windows size is 25 millisecond, number of filter bank channels is 26, cepstral liftering coefficient is 22 with 12 cepstral parameters and the feature vector is normalized. K-means is applied to the MFCC coefficients of all the training data to generate a code book of 64 elements. Finally, each data is converted to a sequence of histograms. We use the same set of parameters as the previous experiment.

The experimental results are reported in Tab. 6.3. From the table, we can find that the improved method consistently outperforms than both plain HMM and also the baseline method in all six dimensions. We also find that the baseline method gets low accuracy on this experiment, which can be explained by that the length of the audio (in number of temporal frames) varies dramatically and the baseline method obviously cannot handle this variation very well.

6.7 Discussions and Conclusions

In this chapter, we presented a new formulation for the problem of learning temporal models using only relative information. Algorithms were developed under the formulation, and experiments using both synthetic and real data were performed to verify the performance of the proposed method. In essence, the proposed method attempts to learn an HMM with relative constraints. Such a setting is useful for many practical applications where relative attributes are easier to obtain while explicit labeling is difficult to get. The application of video-based surgical training was the focus of this study, and the evaluation results using realistic data suggests that the proposed method provides a promising solution to the problem of motion skill evaluation from videos. For future work, we plan to extend the proposed method to cover different observation models so that more types of applications may be handled.

Chapter 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

In this dissertation, we presented our study on problems of how to model the semantic information of the visual data and incorporate it into the sparse learning methods. In this work, we identified four problems which are of great importance and broad interest to the community. Specifically, a novel approach was proposed to incorporate label information to learn a dictionary which is not only reconstructive but also discriminative; considering the formation process of face images, a novel image decomposition approach for an ensemble of correlated images was proposed, where a subspace is built from the decomposition and applied to face recognition; based on the observation that, the foreground (or salient) objects are sparse in input domain and the background is sparse in frequency domain, a novel and efficient spatio-temporal saliency detection algorithm was proposed to identify the salient regions in video; and a novel hidden Markov model learning approach was proposed by utilizing the pairwise comparisons among the data, which is easier to obtain and more meaningful, consistent than tradition labels, in many scenarios, e.g., evaluating motion skills in surgical simulations.

In those four problems, different types of semantic information were modeled and incorporated in designing sparse learning algorithms for the corresponding visual computing tasks. Several real world applications were selected to demonstrate the effectiveness of the proposed methods, including, face recognition, saliency detection in video, abnormality detection, motion analysis and emotion recognition. In those applications, data of different modalities were involved, ranging from audio signal, image to video. Experiments on large scale real world data with comparisons to

state-of-art methods confirmed the proposed approaches deliver manifest advantages, showing adding that semantic information dramatically improve the performances of the general sparse learning methods.

The major work has been published or submitted to peer-reviewed venues: the work in Chapter 3 was initially reported in Zhang and Li (2010a); the work in Chapter 4 has been reported in Zhang and Li (2010b) Zhang and Li (2012) and part of the work in Chapter 6 was reported in Zhang and Li (2013).

7.2 Future Work

The current work may potentially be further improved from the following aspects. In D-KSVD, we are currently using square loss for measuring the classification error. Though achieving good results in face recognition tasks, it can still be improved by using more sophisticated loss type, e.g., logistic loss or hinge loss.

As regarding to JSM-MC, one potential direction could be removing the assumption that the low-rank conditions of the training images are known and the constraint that each image is only related to one imaging condition. To this end, we plan to expand the current algorithm by incorporating another step that attempts to estimate the coefficient of assigning a condition to each image, during the optimization iteration. This problem can be formulated as following:

$$\begin{aligned} \mathbb{C}, \mathbb{A}, \mathbb{E} \quad &= \quad \underset{\mathbb{C},\mathbb{A},\mathbb{E}}{\operatorname{argmin}} \sum_{i,j} \|\mathbf{A}_j\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}\|_1 \\ &s.t. \quad \mathbf{X}_{i,j} = \mathbf{C}_i + \sum_k \mathbf{A}_k \alpha_{i,j}(k) + \mathbf{E}_{i,j}, \ \forall \mathbf{X}_{i,j} \in \mathbb{X} \\ &\qquad \|\alpha_{i,j}\|_1 \leq \tau, \ \forall (i,j) \in \Omega \end{aligned} \tag{7.1}$$

where $\alpha_{i,j}$ is a vector indicating the condition of $\mathbf{X}_{i,j}$, i.e., $\mathbf{X}_{i,j}$ takes Condition k, if and only if $\alpha_{i,j}(k) \neq 0$. This formulation is more general than the original formulation however also involves a more complex optimization algorithm, which shall be the direction of future work.

Chapter 5 demonstrates the effectiveness of the proposed spatiotemporal saliency in detecting salient region from the video. However, more efforts are needed on the theoretic analysis of the proposed method. e.g., some psychology studies may help us to reveal why the proposed method is able to identify the salient volumes.

Finally, being effective in modeling the motion skills in surgical simulation, the relative HMM described in this dissertation still have several limitations, e.g., only multinomial observation model is allowed; and the current algorithm can only find a local optimum. Thus for the future work, we plan to extend the algorithm to enable other observation model, e.g., Gaussian mixture model; and improve the current algorithm by using simulated annealing for finding better local optimum.

ACKNOWLEDGEMENTS *The work was supported in part by a grant (Grant No. 0904778) from the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.*

Adam, A., E. Rivlin, I. Shimshoni and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors", PAMI **30**, 3, 555 –560 (2008).

Aharon, M., M. Elad and A. Bruckstein, "K-SVD: Design of dictionaries for sparse representation", Proceedings of SPARS **5** (2005).

Aharon, M., M. Elad and A. Bruckstein, "¡ img src=", Signal Processing, IEEE Transactions on **54**, 11, 4311–4322 (2006).

Alexe, B., T. Deselaers and V. Ferrari, "Measuring the objectness of image windows", PAMI **34**, 11, 2189 –2202 (2012).

Altun, Y., I. Tsochantaridis, T. Hofmann *et al.*, "Hidden markov support vector machines", in "MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-", vol. 20, p. 3 (2003).

Ban, S., I. Lee and M. Lee, "Dynamic visual selective attention model", Neurocomputing **71**, 4, 853–856 (2008).

Basri, R. and D. W. Jacobs, "Lambertian reflectance and linear subspaces", Pattern Analysis and Machine Intelligence, IEEE Transactions on **25**, 2, 218–233 (2003).

Baum, L. E., T. Petrie, G. Soules and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains", The Annals of Mathematical Statistics **41**, 1, pp. 164–171, URL `http://www.jstor.org/stable/2239727` (1970).

Belhumeur, P. N., J. P. Hespanha and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection", Pattern Analysis and Machine Intelligence, IEEE Transactions on **19**, 7, 711–720 (1997).

Bengio, S., F. Pereira, Y. Singer and D. Strelow, "Group sparse coding", Adv. NIPS (2009).

Bertsekas, D. P., "Constrained optimization and lagrange multiplier methods", Computer Science and Applied Mathematics, Boston: Academic Press, 1982 **1** (1982).

Bian, P. and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency", NIPS pp. 251–258 (2009).

Borji, A. and L. Itti, "State-of-the-art in visual attention modeling", PAMI **PP**, 99, 1 (2012).

Borji, A., D. N. Sihite and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study", (2012).

Bruce, N. and J. Tsotsos, "Saliency based on information maximization", in "Advances in neural information processing systems", pp. 155–162 (2005).

Bryt, O. and M. Elad, "Compression of facial images using the k-svd algorithm", Journal of Visual Communication and Image Representation **19**, 4, 270–282 (2008a).

Bryt, O. and M. Elad, "Improving the k-svd facial image compression using a linear deblocking method", in "Electrical and Electronics Engineers in Israel, 2008. IEEEI 2008. IEEE 25th Convention of", pp. 533–537 (IEEE, 2008b).

Cai, J. F., E. J. Candes and Z. Shen, "A singular value thresholding algorithm for matrix completion", preprint (2008).

Candes, E. and Y. Plan, "Matrix completion with noise", Proceedings of the IEEE URL `http://arxiv.org/pdf/0903.3131` (2009).

Candes, E. and J. Romberg, "l1-magic: Recovery of sparse signals via convex programming", URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf **4** (2005).

Candès, E. J., J. Romberg and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information", Information Theory, IEEE Transactions on **52**, 2, 489–509 (2006).

Chang, C.-C. and C.-J. Lin, "LIBSVM: A library for support vector machines", ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm` (2011).

Chen, S., C. Cowan and P. Grant, "Orthogonal least squares learning algorithm for radial basis function networks", Neural Networks, IEEE Transactions on **2**, 2, 302–309 (1991).

Collins, M., "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms", in "Proceedings of the ACL 2002", vol. 10, pp. 1–8 (Association for Computational Linguistics, 2002).

Cong, Y., J. Yuan and J. Liu, "Sparse reconstruction cost for abnormal event detection", in "CVPR 2011", pp. 3449 –3456 (2011).

Dollar, P., V. Rabaud, G. Cottrell and S. Belongie, "Behavior recognition via sparse spatio-temporal features", in "Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on", pp. 65 – 72 (2005).

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

Duan, F., Y. Zhang, N. Pongthanya, K. Watanabe, H. Yokoi and T. Arai, "Analyzing human skill through control trajectories and motion capture data", in "Automation Science and Engineering, 2008. CASE 2008. IEEE International Conference on", pp. 454 –459 (2008).

El Ayadi, M., M. S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition **44**, 3, 572–587 (2011).

Elad, M. and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries", Image Processing, IEEE Transactions on **15**, 12, 3736–3745 (2006).

Felzenszwalb, P., R. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models", Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**, 9, 1627–1645 (2010).

Figueiredo, M. A., R. D. Nowak and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems", Selected Topics in Signal Processing, IEEE Journal of **1**, 4, 586–597 (2007).

Fox, E., *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*, Ph.D. thesis, MIT, Cambridge, MA (2009).

Gao, D., S. Han and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition", PAMI **31**, 6, 989 –1005 (2009).

Garcia-Diaz, A., X. R. Fdez-Vidal, X. M. Pardo and R. Dosil, "Decorrelation and distinctiveness provide with human-like saliency", in "Advanced Concepts for Intelligent Vision Systems", pp. 343–354 (Springer, 2009).

Georghiades, A., P. Belhumeur and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose", IEEE Transactions on Pattern Analysis and Machine Intelligence **23**, 6, 643–660 (2001).

Goldfarb, D. and S. Ma, "Convergence of fixed-point continuation algorithms for matrix rank minimization", Foundations of Computational Mathematics pp. 1–28 (2011).

Gorelick, L., M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes", PAMI **29**, 12, 2247–2253 (2007).

Guha, T. and R. K. Ward, "Learning sparse representations for human action recognition", Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**, 8, 1576–1588 (2012).

Guo, C., Q. Ma and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform", in "CVPR 2008", pp. 1 –8 (2008).

Guo, C. and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression", Image Processing, IEEE Transactions on **19**, 1, 185 –198 (2010).

Hale, E. T., W. Yin and Y. Zhang, "Fixed-point continuation for ℓ_1-minimization: Methodology and convergence", SIAM Journal on Optimization **19**, 3, 1107–1130, URL `http://link.aip.org/link/?SJE/19/1107/1` (2008).

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

Hou, X., J. Harel and C. Koch, "Image signature: Highlighting sparse salient regions", PAMI **34**, 1, 194–201 (2012).

Hou, X. and L. Zhang, "Saliency detection: A spectral residual approach", in "CVPR 2007", pp. 1 –8 (2007).

Hou, X. and L. Zhang, "Dynamic visual attention: Searching for coding length increments", NIPS **21**, 681–688 (2008).

Huber, D. E. and C. G. Healey, "Visualizing data with motion", in "Visualization, 2005. VIS 05. IEEE", pp. 527–534 (IEEE, 2005).

Itti, L. and P. Baldi, "Bayesian surprise attracts human attention", NIPS **18**, 547 (2006).

Itti, L., N. Dhavale and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention", in "Optical Science and Technology, SPIE's 48th Annual Meeting", pp. 64–78 (International Society for Optics and Photonics, 2004).

Itti, L., C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", PAMI **20**, 11, 1254 –1259 (1998).

Itti, R., Laurent; Carmi, "Eye-tracking data from human volunteers watching complex video stimuli", Online, URL `CRCNS.org` (2009).

Juang, B. and L. Rabiner, "The segmental¡ e1¿ k¡/e1¿-means algorithm for estimating parameters of hidden markov models", Acoustics, Speech and Signal Processing, IEEE Transactions on **38**, 9, 1639–1641 (1990).

Judd, T., K. Ehinger, F. Durand and A. Torralba, "Learning to predict where humans look", in "ICCV 2009", pp. 2106 –2113 (2009).

Kadar, I. and O. Ben-Shahar, "Small sample scene categorization from perceptual relations", in "CVPR 2012", pp. 2711 –2718 (2012).

Kahol, K., N. C. Krishnan, V. N. Balasubramanian, S. Panchanathan, M. Smith and J. Ferrara, "Measuring movement expertise in surgical tasks", in "Proceedings of the 14th annual ACM international conference on Multimedia", MULTIMEDIA '06, pp. 719–722 (ACM, New York, NY, USA, 2006).

Kanan, C. and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics", in "Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on", pp. 2472 –2479 (2010).

Kim, J. and K. Grauman, "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates", in "CVPR 2009", pp. 2921 –2928 (2009).

Kim, K. H., S. Bang and S. Kim, "Emotion recognition system using short-term monitoring of physiological signals", Medical and biological engineering and computing **42**, 3, 419–427 (2004).

Kläser, A., *Learning human actions in video*, Ph.D. thesis, Université de Grenoble, URL `http://lear.inrialpes.fr/pubs/2010/Kla10` (2010).

Knyazev, A. and M. Argentati, "Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates", SIAM Journal on Scientific Computing **23**, 6, 2008–2040 (2002).

Kovashka, A., D. Parikh and K. Grauman, "Whittlesearch: Image search with relative attribute feedback", in "CVPR 2012", pp. 2973 –2980 (2012).

Kumar, N., A. Berg, P. Belhumeur and S. Nayar, "Attribute and simile classifiers for face verification", in "ICCV 2009", pp. 365 –372 (2009a).

Kumar, R., A. Banerjee and B. Vemuri, "Volterrafaces: Discriminant analysis using volterra kernels", in "Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on", pp. 150 –155 (2009b).

Laptev, I., "On space-time interest points", IJCV **64**, 2, 107–123 (2005).

Lee, K., J. Ho and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting", IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 684–698 (2005).

Li, Q., Y. Zhou and J. Yang, "Saliency based image segmentation", in "Multimedia Technology (ICMT), 2011 International Conference on", pp. 5068 –5071 (2011).

Li, Y., Y. Zhou, J. Yan, Z. Niu and J. Yang, "Visual saliency based on conditional entropy", ACCV 2009 pp. 246–257 (2010).

Lin, Z., M. Chen, L. Wu and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices", Arxiv preprint arXiv:1009.5055 (2010).

Lin, Z., A. Ganesh, J. Wright, L. Wu, M. Chen and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix", Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) (2009).

Liu, B., J. Huang, L. Yang and C. Kulikowsk, "Robust tracking using local sparse appearance model and k-selection", in "Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on", pp. 1313–1320 (IEEE, 2011).

Liu, J., S. Chen and X. Tan, "Fractional order singular value decomposition representation for face recognition", Pattern Recogn. **41**, 378–395, URL `http://dl.acm.org/citation.cfm?id=1284917.1285188` (2008).

Ma, Y.-F., X.-S. Hua, L. Lu and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization", Multimedia, IEEE Transactions on **7**, 5, 907 – 919 (2005).

Mahadevan, V., W. Li, V. Bhalodia and N. Vasconcelos, "Anomaly detection in crowded scenes", in "CVPR 2010", pp. 1975 –1981 (2010).

Mahadevan, V. and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes", PAMI **32**, 1, 171–177 (2010).

Mairal, J., F. Bach, J. Ponce, G. Sapiro and A. Zisserman, "Discriminative learned dictionaries for local image analysis", in "Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on", pp. 1–8 (IEEE, 2008a).

Mairal, J., F. Bach, J. Ponce, G. Sapiro, A. Zisserman *et al.*, "Supervised dictionary learning", (2008b).

Mairal, J., M. Elad and G. Sapiro, "Sparse representation for color image restoration", Image Processing, IEEE Transactions on **17**, 1, 53–69 (2008c).

Mancas, M., *Computational attention: Modelisation and application to audio and image processing*, Ph.D. thesis, PhD. Thesis, University of Mons (2007).

Marat, S., T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos", IJCV **82**, 3, 231–243 (2009).

Martinez, A. and R. Benavente, "The AR face database", Tech. rep., CVC Technical report (1998).

Martinez, A. and R. Benavente, "The ar face database, 1998", Computer Vision Center, Technical Report **3** (2007).

Mehran, R., A. Oyama and M. Shah, "Abnormal crowd behavior detection using social force model", in "CVPR 2009", pp. 935 –942 (2009).

Merhav, N. and Y. Ephraim, "Maximum likelihood hidden markov modeling using a dominant sequence of states", Signal Processing, IEEE Transactions on **39**, 9, 2111 –2115 (1991).

Mital, P. K., T. J. Smith, R. L. Hill and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion", Cognitive Computation **3**, 1, 5–24 (2011).

Mori, H., T. Satake, M. Nakamura and H. Kasuya, "Uu database: A spoken dialogue corpus for studies on paralinguistic information in expressive conversation", pp. 427–434, URL `http://uudb.speech-lab.org/` (2008).

Nagesh, P. and B. Li, "A compressive sensing approach for expression-invariant face recognition", in "Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on", pp. 1518 –1525 (2009).

Nwe, T. L., S. W. Foo and L. C. De Silva, "Speech emotion recognition using hidden markov models", Speech communication **41**, 4, 603–623 (2003).

Ölveczky, B., S. Baccus and M. Meister, "Segregation of object and background motion in the retina", Nature **423**, 6938, 401–408 (2003).

Parikh, D. and K. Grauman, "Relative attributes", in "ICCV 2011", pp. 503 –510 (2011).

Parikh, D., A. Kovashka, A. Parkash and K. Grauman, "Relative attributes for enhanced human-machine communication", in "Twenty-Sixth AAAI", (2012).

Pham, D.-S. and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition", in "Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on", pp. 1–8 (IEEE, 2008).

Portilla, J. and E. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients", International Journal of Computer Vision **40**, 1, 49–70 (2000).

Raghavendra, R., A. Del Bue, M. Cristani and V. Murino, "Optimizing interaction force for global anomaly detection in crowded scenes", in "Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on", pp. 136 –143 (2011).

Reddy, V., C. Sanderson and B. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture", in "Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on", pp. 55 –61 (2011).

Rodriguez, M., J. Ahmed and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition", in "CVPR 2008", pp. 1–8 (2008).

Rosen, J., M. Solazzo, B. Hannaford and M. Sinanan, "Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden markov model", Computer Aided Surgery **7**, 1, 49–61 (2002).

Satoshi, S. and H. Fumio, "Skill evaluation from observation of discrete hand movements during console operation", Journal of Robotics **2010** (2010).

Schuldt, C., I. Laptev and B. Caputo, "Recognizing human actions: a local svm approach", in "ICPR 2004", vol. 3, pp. 32 – 36 Vol.3 (2004).

Schuller, B., G. Rigoll and M. Lang, "Hidden markov model-based speech emotion recognition", in "Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on", vol. 2, pp. II–1 (IEEE, 2003).

Schultz, M. and T. Joachims, "Learning a distance metric from relative comparisons", NIPS p. 41 (2004).

Seo, H. J. and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance", Journal of Vision **9**, 12, URL `http://www.journalofvision.org/content/9/12/15.abstract` (2009).

Sharma, G., F. Jurie and C. Schmid, "Discriminative spatial saliency for image classification", in "CVPR 2010", pp. 3506 –3513 (2012).

Sharon, Y., J. Wright and Y. Ma, "Computation and relaxation of conditions for equivalence between l1 and l0 minimization", submitted to IEEE Transactions on Information Theory **5** (2007).

Sim, T., S. Baker and M. Bsat, "The CMU pose, illumination, and expression (PIE) database", in "Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition", (2002).

Simoncelli, E. P. and O. Schwartz, "Modeling surround suppression in v1 neurons with a statistically-derived normalization model", (1999).

Sloin, A. and D. Burshtein, "Support vector machine training for improved hidden markov modeling", Signal Processing, IEEE Transactions on **56**, 1, 172 –188 (2008).

Suzuki, S., N. Tomomatsu, F. Harashima and K. Furuta, "Skill evaluation based on state-transition model for human adaptive mechatronics (ham)", in "Industrial Electronics Society, 2004. IECON 2004. 30th Annual Conference of IEEE", vol. 1, pp. 641–646 (IEEE, 2004).

Tarasov, A. and S. J. Delany, "Benchmarking classification models for emotion recognition in natural speech: a multi-corporal study", in "Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on", pp. 841–846 (IEEE, 2011).

Torralba, A., "Modeling global scene factors in attention", JOSA A **20**, 7, 1407–1418 (2003).

Tropp, J. A., "Just relax: Convex programming methods for identifying sparse signals in noise", Information Theory, IEEE Transactions on **52**, 3, 1030–1051 (2006).

Turk, M. A. and A. P. Pentland, "Face recognition using eigenfaces", in "Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on", pp. 586–591 (IEEE, 1991).

Wang, G., D. Forsyth and D. Hoiem, "Comparative object similarity for improved recognition with few or no examples", in "CVPR 2010", pp. 3525–3532 (IEEE, 2010).

Watanabe, K. and M. Hokari, "Kinematical analysis and measurement of sports form", Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on **36**, 3, 549–557 (2006).

Willems, G., T. Tuytelaars and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector", ECCV 2008 pp. 650–663 (2008).

Woodland, P. C., J. J. Odell, V. Valtchev and S. J. Young, "Large vocabulary continuous speech recognition using htk", in "Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on", vol. 2, pp. II–125 (IEEE, 1994), URL `http://htk.eng.cam.ac.uk/`.

Wright, J., A. Ganesh, S. Rao, Y. Peng and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization", in "Proc. of Neural Information Processing Systems", vol. 3 (2009a).

Wright, J., A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, "Robust face recognition via sparse representation", Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**, 2, 210–227 (2009b).

Wu, S., B. Moore and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes", in "CVPR 2010", pp. 2054 –2060 (2010).

Yan, J., M. Zhu, H. Liu and Y. Liu, "Visual saliency detection via sparsity pursuit", Signal Processing Letters, IEEE **17**, 8, 739–742 (2010).

Yang, J., Z. Wang, Z. Lin, S. Cohen and T. Huang, "Coupled dictionary training for image super-resolution", Image Processing, IEEE Transactions on **21**, 8, 3467–3478 (2012).

Yang, J., K. Yu, Y. Gong and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification", in "Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on", pp. 1794–1801 (IEEE, 2009).

Zhai, Y. and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues", in "Proceedings of the 14th ACM Multimedia", MULTIMEDIA '06, pp. 815–824 (ACM, New York, NY, USA, 2006), URL http://doi.acm.org/10.1145/1180639.1180824.

Zhang, L., M. Tong and G. Cottrell, "Sunday: Saliency using natural statistics for dynamic analysis of scenes", in "Proceedings of the 31st Annual Cognitive Science Conference", pp. 2944–2949 (AAAI Press Cambridge, MA, 2009).

Zhang, Q. and B. Li, "Discriminative K-SVD for dictionary learning in face recognition", in "Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on", pp. 2691–2698 (IEEE, 2010a).

Zhang, Q. and B. Li, "JOINT SPARSITY MODEL WITH MATRIX COMPLETION FOR AN ENSEMBLE OF FACE IMAGES", in "Image Processing (ICIP), 2010 17th IEEE International Conference on", p. preprint (IEEE, Hongkong, 2010b).

Zhang, Q. and B. Li, "Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model", in "Proceedings of the 2011 international ACM workshop on Medical multimedia analysis and retrieval", MMAR '11, pp. 19–24 (ACM, New York, NY, USA, 2011).

Zhang, Q. and B. Li, "Mining discriminative components with low-rank and sparsity constraints for face recognition", in "Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 1469–1477 (ACM, 2012).

Zhang, Q. and B. Li, "Relative hidden markov models for evaluating motion skills", in "CVPR 2013", pp. ?–? (IEEE, 2013).

Zhao, W., R. Chellappa, P. Phillips and A. Rosenfeld, "Face recognition: A literature survey", Acm Computing Surveys (CSUR) **35**, 4, 399–458 (2003).

Zhao, Z., G.-J. Ahn, J.-J. Seo and H. Hu, "On the security of picture gesture authentication", in "22nd USENIX Security Symposium (Security)", (USENIX, 2013).

Zhaoping, L. and P. Dayan, "Pre-attentive visual selection", Neural Networks **19**, 9, 1437–1439 (2006).

Zhou, M., H. Chen, L. Ren, G. Sapiro, L. Carin and J. W. Paisley, "Non-parametric bayesian dictionary learning for sparse image representations", in "Advances in neural information processing systems", pp. 2295–2303 (2009).

APPENDIX A

PROOF OF THEOREM 1 IN CHAPTER 3

Proposition 1 The sequences of $\tilde{Y}_{i,j}^{t+1}$, $\sum_i \hat{Y}_{i,j}^{t+1}$, $\sum_j \mathbf{Y}_{i,j}^{t+1}$ and $\dot{Y}_{i,j}^{t+1}$ are all bounded $\forall i, j$, where

$$\begin{aligned}
\mathbf{Y}_{i,j}^{t+1} &= \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^{t+1} - \mathbf{A}_j^{t+1} - \mathbf{E}_{i,j}^{t+1}) + \mathbf{Y}_{i,j}^t \\
\hat{Y}_{i,j}^{t+1} &= \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^{t} - \mathbf{A}_j^{t+1} - \mathbf{E}_{i,j}^{t+1} + \mathbf{Y}_{i,j}^t \\
\tilde{Y}_{i,j}^{t+1} &= \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^{t} - \mathbf{A}_j^{t} - \mathbf{E}_{i,j}^{t+1}) + \mathbf{Y}_{i,j}^t \\
\dot{Y}_{i,j}^{t+1} &= \mu_{i,j}^t(\mathbf{X}_{i,j} - \dot{C}_i^{t+1} - \dot{A}_j^{t+1} - \dot{E}_{i,j}^{t+1}) + \dot{Y}_{i,j}^t
\end{aligned}$$

and $(\dot{C}^{t+1}, \dot{A}^{t+1}, \dot{E}^{t+1})$ is the optimal solution to the problem $\min_{\mathbb{C},\mathbb{A},\mathbb{E}} L(\mathbb{C}, \mathbb{A}, \mathbb{E}, \dot{Y}^t, \mu^t)$ with $\dot{Y}^t = \{\dot{Y}_{i,j}^t\}_{i,j=1}^{N,M}$.

Proof Let's write the Lagrange function in 6.11 as:

$$\begin{aligned}
& L(\{\mathbf{C}_i^t\}_i, \{\mathbf{A}_j^t\}_j, \{\mathbf{E}_{i,j}^t\}_{i,j}, \{\mathbf{Y}_{i,j}^t\}_{i,j}, \{\mu^t\}_{i,j}) \\
= & \sum_{i,j} \|\mathbf{A}_j^t\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}^t\|_1 \\
+ & \frac{\mu_{i,j}^t}{2}\|\mathbf{X}_{i,j} - \mathbf{C}_i^t - \mathbf{A}_j^t - \mathbf{E}_{i,j}^t\|_F^2 \\
+ & < \mathbf{Y}_{i,j}^t, \mathbf{X}_{i,j} - \mathbf{C}_i^t - \mathbf{A}_j^t - \mathbf{E}_{i,j}^t >
\end{aligned}$$

For simplicity, we will use $L(\mathbb{C}^t, \mathbb{A}^t, \mathbb{E}^{t+1}, \mathbb{Y}^t, \mu^t)$ instead of $L(\{\mathbf{C}_i^t\}_i, \{\mathbf{A}_j^t\}_j, \{\mathbf{E}_{i,j}^{t+1}\}_{i,j}, \{\mathbf{Y}_{i,j}^t\}_{i,j}, \{\mu^t\}_{i,j})$. The subgradient of $L(\mathbb{C}^t, \mathbb{A}^t, \mathbb{E}, \mathbb{Y}^t, \mu_t)$ over $\mathbf{E}_{i,j}$ is

$$\lambda_{i,j}\partial\|\mathbf{E}_{i,j}\|_1 - \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^t - \mathbf{A}_j^t - \mathbf{E}_{i,j}) - \mathbf{Y}_{i,j}^t$$

As $\mathbf{E}_{i,j}^{t+1}$ is optimal for the problem $\underset{\mathbf{E}_{i,j}}{\operatorname{argmin}} L(\mathbb{C}^t, \mathbb{A}^t, \mathbb{E}, \mathbb{Y}^t, \mu_t)$

$$0 \in \lambda_{i,j}\partial\|\mathbf{E}_{i,j}^t\|_1 - \tilde{Y}_{i,j}^{t+1}$$

i.e., $\tilde{Y}_{i,j}^{t+1} \in \lambda_{i,j}\|\mathbf{E}_{i,j}^{t+1}\|_1$; and according to the Theorem 3 of Lin *et al.* (2010), $\tilde{Y}_{i,j}^{t+1}$ is bounded $\forall i, j$. Similarly, we can also show that $\sum_i \hat{Y}_{i,j}^{t+1}$, $\sum_j \mathbf{Y}_{i,j}^{t+1}$ and $\dot{Y}_{i,j}^{t+1}$ are bounded $\forall i, j$.

Proposition 2 The sequences of $(\mathbb{C}^{t+1}, \mathbb{A}^{t+1}, \mathbb{E}^{t+1})$ is bounded.

Proof For Algorithm 1, we can find that:

$$\begin{aligned}
L(\mathbb{C}^{t+1}, \mathbb{A}^{t+1}, \mathbb{E}^{t+1}, \mathbb{Y}^t, \mu^t) &\leq L(\mathbb{C}^t, \mathbb{A}^{t+1}, \mathbb{E}^{t+1}, \mathbb{Y}^t, \mu^t) \\
\leq L(\mathbb{C}^t, \mathbb{A}^t, \mathbb{E}^{t+1}, \mathbb{Y}^t, \mu^t) &\leq L(\mathbb{C}^t, \mathbb{A}^t, \mathbb{E}^t, \mathbb{Y}^t, \mu^t) \\
= L(\mathbb{C}^t, \mathbb{A}^t, \mathbb{E}^t, \mathbb{Y}^{t-1}, \mu^{t-1}) &+ \sum_{i,j} \frac{\mu_{i,j}^{t-1} + \mu_{i,j}^t}{(\mu_{i,j}^t)^2}\|\mathbf{Y}_{i,j}^t - \mathbf{Y}_{i,j}^{t-1}\|_F^2
\end{aligned}$$

By boundedness of assumption that $\sum_{t=1}^{\infty} \mu_{i,j}^{t+1}(\mu_{i,j}^t)^{-2} < \infty$ and $\sum_j \mathbf{Y}_{i,j}^t$ $\forall i, j$, we have $L(\mathbb{C}^{t+1}, \mathbb{A}^{t+1}, \mathbb{E}^{t+1}, \mathbb{Y}^t, \mu^t)$ is upper bounded. Thus $\sum_{i,j}\|\mathbf{A}_j^t\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}^t\|_1$ is bounded.

Proposition 3 The accumulation point $(\dot{C}^*, \dot{A}^*, \dot{E}^*)$ for sequences $(\dot{C}^{t+1}, \dot{A}^{t+1}, \dot{E}^{t+1})$ is optimal for the problem in Eqn. 4.5.

Proof For $(\dot{C}^{t+1}, \dot{A}^{t+1}, \dot{E}^{t+1})$, we have the following:

$$\begin{aligned} & L(\dot{C}^{t+1}, \dot{A}^{t+1}, \dot{E}^{t+1}, \dot{Y}^t, \mu^t) = \min_{\mathbb{C},\mathbb{A},\mathbb{E}} L(\mathbb{C}, \mathbb{A}, \mathbb{E}, \dot{Y}^t, \mu^t) \\ \leq & \min_{\mathbf{C}_i+\mathbf{A}_j+\mathbf{E}_{i,j}=\mathbf{X}_{i,j},\forall(i,j)} L(\mathbb{C}, \mathbb{A}, \mathbb{E}, \dot{Y}^t, \mu^t) \\ \leq & \min_{\mathbf{C}_i+\mathbf{A}_j+\mathbf{E}_{i,j}=\mathbf{X}_{i,j},\forall(i,j)} \sum_{i,j} \|\mathbf{A}_j\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}\|_1 \\ = & f^* \end{aligned}$$

We also have:

$$\begin{aligned} & \sum_{i,j} \|\dot{A}_j^{t+1}\|_* + \lambda_{i,j}\|\dot{E}_{i,j}^{t+1}\|_1 \\ = & L(\dot{C}^{t+1}, \dot{A}^{t+1}, \dot{E}^{t+1}, \dot{Y}^t, \mu^t) - \sum_{i,j} \frac{\|\dot{Y}_{i,j}^t - \dot{Y}_{i,j}^{t-1}\|_F^2}{2\mu_{i,j}^t} \\ \leq & f^* - \sum_{i,j} \frac{\|\dot{Y}_{i,j}^t - \dot{Y}_{i,j}^{t-1}\|_F^2}{2\mu_{i,j}^t} = f^* + O(\sum_{i,j} (\mu_{i,j}^t)^{-1}) \end{aligned}$$

where we use the knowledge that $\dot{Y}_{i,j}^{t+1}$ is bounded $\forall i, j$. Take $t \to \infty$, we have $\sum_{i,j} \|\dot{A}_j^*\|_* + \lambda_{i,j}\|\dot{E}_{i,j}^*\|_1 = f^*$. Using $(\dot{Y}_{i,j}^t - \dot{Y}_{i,j}^{t-1}) = \mu_{i,j}^{t-1}(\dot{X}_{i,j} - \dot{C}_i^{t-1} - \dot{A}_j^{t-1} - \dot{E}_{i,j}^{t-1})$ and boundedness of $\dot{Y}_{i,j}^{t+1}$ $\forall i, j$, we also have $\mathbf{X}_{i,j} - \dot{C}_i^* - \dot{A}_j^* - \dot{E}_{i,j}^* = 0$ $\forall i, j$. Thus $(\dot{C}^*, \dot{A}^*, \dot{E}^*)$ is the optimal solution for Eqn. 4.5.

By $\mathbf{X}_{i,j} - \mathbf{C}_i^{t+1} - \mathbf{A}_j^{t+1} - \mathbf{E}_{i,j}^{t+1} = \mu_{i,j}^t(\mathbf{Y}_{i,j}^{t+1} - \mathbf{Y}_{i,j}^t)$ and boundedness of $\mathbf{Y}_{i,j}^t$, we have $\lim_{t\to\infty} \mathbf{C}_i^{t+1} + \mathbf{A}_j^{t+1} + \mathbf{E}_{i,j}^{t+1} = \mathbf{X}_{i,j}$ $\forall i, j$, i.e., $(\mathbb{C}^{t+1}, \mathbb{A}^{t+1}, \mathbb{E}^{t+1})$ approaches to a feasible solution. In addition, we have

$$\|\sum_i \mathbf{A}_j^{t+1} - \mathbf{A}_j^t\|_F = \|\sum_i (\mu_{i,j}^t)^{-1}(\hat{Y}_{i,j}^{t+1} - \tilde{Y}_{i,j}^{t+1})\|_F$$

With the assumption $\sum_{t=1}^{\infty} (\mu_{i,j}^t)^{-1} < \infty$, boundedness of $\sum_i \hat{Y}_{i,j}^{t+1}$ and $\tilde{Y}_{i,j}^t$, $\mathbf{A}_j^{t+1}$ has a limit $\mathbf{A}_j^*$. Similarly:

$$\|\sum_j \mathbf{A}_j^{t+1} - \mathbf{A}_j^t + \mathbf{C}_i^{t+1} - \mathbf{C}_i^t\|_F = \|\sum_j (\mu_{i,j}^t)^{-1}(\mathbf{Y}_{i,j}^{t+1} - \tilde{Y}_{i,j}^{t+1})\|_F$$

Thus $\lim_{t\to\infty} \sum_j \mathbf{A}_j^{t+1} - \mathbf{A}_j^t + \mathbf{C}_i^{t+1} - \mathbf{C}_i^t = 0$. Since $\mathbf{A}_j^{t+1}$ has limit $\mathbf{A}_j^*$, then $\mathbf{C}_i^{t+1}$ has limit $\mathbf{C}_i^*$, then $\mathbf{E}_{i,j}^{t+1}$ has limit $\mathbf{X}_{i,j} - \mathbf{A}_j^* - \mathbf{C}_i^*$. So $(\mathbb{C}^*, \mathbb{A}^*, \mathbb{E}^*)$ is a feasible solution.

Considering the subgradients and the optimality of $\mathbf{E}_{i,j}^{t+1}$ and $\mathbf{A}_j^{t+1}$, we have $\tilde{Y}_{i,j}^{t+1} \in \partial\|\mathbf{E}_{i,j}^{t+1}\|_1$ and $\sum_i \hat{Y}_{i,j}^{t+1} \in \partial\|\mathbf{A}_j^{t+1}\|_*$. According to the property of subgradients:

$$\begin{aligned} & (\sum_{i,j} \|\mathbf{A}_j^{t+1}\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}^{t+1}\|_1) - (\sum_{i,j} \|\dot{A}_j^{t+1}\|_* + \lambda_{i,j}\|\dot{E}_{i,j}^{t+1}\|_1) \\ \leq & \sum_{i,j} - < \hat{Y}_{i,j}^{t+1}, \dot{A}_j^{t+1} - \mathbf{A}_j^{t+1} > - < \tilde{Y}_{i,j}^{t+1}, \dot{E}_{i,j}^{t+1} - \mathbf{E}_{i,j}^{t+1} > \\ = & \sum_{i,j} -\mu_{i,j}^t < \mathbf{E}_{i,j}^{t+1} - \mathbf{E}_{i,j}^t, \dot{A}_j^{t+1} - \mathbf{A}_j^{t+1} > \\ - & \frac{< \tilde{Y}_{i,j}^{t+1}, \tilde{Y}_{i,j}^{t+1} - \tilde{Y}_{i,j}^t >}{\mu_{i,j}^t} - \frac{< \tilde{Y}_{i,j}^{t+1}, \dot{Y}_{i,j}^{t+1} - \dot{Y}_{i,j}^t >}{\mu_{i,j}^t} \end{aligned}$$

By Proposition 1 and 2 that $\mathbf{Y}_{i,j}^{t+1}$, $\dot{Y}_{i,j}^{t+1}$ are bounded; by Proposition 3 that $\sum_{i,j}\|\dot{A}_j^*\|_*+\lambda_{i,j}\|\dot{E}_{i,j}^*\|_1 = f^*$; and by assumption $\lim_{t\to\infty}\mu_{i,j}^t(\mathbf{E}_{i,j}^{t+1}-\mathbf{E}_{i,j}^t)=0$, we have $\sum_{i,j}\|\mathbf{A}_j^*\|_*+\lambda_{i,j}\|\mathbf{E}_{i,j}^*\|_1 = f^*$. That is $(\mathbb{C}^*,\mathbb{A}^*,\mathbb{E}^*)$ is optimal for the problem in Eqn. 4.5. This completes the proof of Theorem 1.

APPENDIX B

COMPARISON OF SALIENCY MAP COMPUTED BY QFT AND FFT

To compare the performances of combining four visual cue via QFT and performances via summation of saliency maps of each visual cues, we design the following experiment. We run 1000 simulations and in each simulation we generate a $r \times c \times 4$ array, where r and c is a random number between $[1, 1000]$ and 4 is the number of feature channels. We compute the saliency map with different methods then measure their similarities via cross-correlation, where 0.91 is reported for QFT and FFT. After smoothing the saliency map with a Gaussian kernel, the correlation is over 0.998. For natural image, we could expect an even higher correlation.

This suggests that, we can compute the saliency map for each visual cue independently and then add them together, which will yield quite similar result by using quaternion Fourier transform. In addition, the proposed method other than QFT provides more flexibility, e.g., we can assign different weights to the visual cues as Judd *et al.* (2009).

APPENDIX C

OBSERVATION MODEL WITH MULTINOMIAL DISTRIBUTION

For multinomial observation model, i.e., $p(\mathbf{X}_t|\phi_{\mathbf{z}_t}) = \prod_{d=1}^{D} \phi_{\mathbf{z}_t}(l)^{\mathbf{X}_t(l)}$, where D is the dimension of each frame, $\mathbf{X}_t(l)$ is the l_{th} dimension of $\mathbf{X}_t$ and $\phi_{\mathbf{z}_t}$ are the parameters of observation model with State $\mathbf{z}_t$, we can further define the following variables for each sequence $\mathbf{X}^i$:

$$\begin{aligned}
\mathbf{n}^i \in \mathbb{R}^{K\times 1} &: \mathbf{n}^i(k) = \delta(\mathbf{z}_1^i = k) \\
\mathbf{O}^i \in \mathbb{R}^{K\times D} &: \mathbf{O}^i(k,d) = \sum_{t:\mathbf{z}_t=k} \mathbf{X}_t^i(d) \\
\mathbf{M}^i \in \mathbb{R}^{K\times K} &: \mathbf{M}^i(k,l) = \sum_{t=2}^{T} \delta(\mathbf{z}_{t-1}^i = k)\delta(\mathbf{z}_t^i = l)
\end{aligned}$$

where $\delta(\cdot)$ is Dirac Delta function. Then the log likelihood with the optimal path can be written as:

$$\begin{aligned}
\log p(\mathbf{X}^i, \mathbf{z}^i|\theta) &= \sum_l \mathbf{n}^i(l) \log \pi(l) \\
&+ \sum_{k,l} \mathbf{M}^i(k,l) \log \mathbf{A}(k,l) \\
&+ \sum_{k,d} \mathbf{O}^i(k,d) \log \phi_k(d) \\
&= \theta^T h(\mathbf{X}^i, \mathbf{z}^i) \qquad \text{(C.1)}
\end{aligned}$$

where $\theta = [\log \pi; \mathrm{vec}(\log \mathbf{A}); \mathrm{vec}(\log \phi)]$, $h(\mathbf{X}^i, \mathbf{z}^i) = [\mathbf{n}^i; \mathrm{vec}(\mathbf{M}^i); \mathrm{vec}(\mathbf{O}^i)]$ and vec converts matrix to vector.

APPENDIX D

RELATED PUBLICATIONS

D.1 CONFERENCE/JOURNAL PUBLICATIONS

- Qiang Zhang and Baoxin Li, Relative Hidden Markov Models for Evaluating Motion Skills, IEEE Computer Vision and Pattern Recognition (CVPR) 2013, Portland, OR

- Lin Chen, Qiongjie Tian, Qiang Zhang and Baoxin Li. Learning Skill-Defining Latent Space in Video-Based Analysis of Surgical Expertise: A Multi-Stream Fusion Approach. NextMed/MMVR20. San Diego, CA, 2013.

- Qiongjie Tian, Lin Chen, Qiang Zhang and Baoxin Li. Enhancing Fundamentals of Laparoscopic Surgery Trainer Box via Designing A Multi-Sensor Feedback System. NextMed/MMVR20. San Diego, CA, 2013.

- Qiang Zhang, Lin Chen, Qiongjie Tian and Baoxin Li. Video-based analysis of motion skills in simulation-based surgical training. SPIE Multimedia Content Access: Algorithms and Systems VII. San Francisco, CA, 2013.

- Qiang Zhang and Baoxin Li. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In Proceedings of the 2011 international ACM workshop on Medical multimedia analysis and retrieval (MMAR '11). ACM [oral], New York, NY, USA, 19-24.

- Zhang, Qiang and Li, Baoxin, Towards Computational Understanding of Skill Levels in Simulation-Based Surgical Training via Automatic Video Analysis, International Symposium on Visual Computing (ISVC) 2010, Las Vegas, NV

- Qiang Zhang and Baoxin Li. Mining Discriminative Components With Low-Rank And Sparsity Constraints for Face Recognition. The 18th ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (SIGKDD 2012).

- Qiang Zhang and Baoxin Li, Joint Sparsity Model with Matrix Completion for an Ensemble of Images, IEEE International Conference on Image Processing (ICIP) 2010, Hong Kong, China

- Qiang Zhang and Baoxin Li, Discriminative K-SVD for Dictionary Learning in Face Recognition, IEEE Computer Vision and Pattern Recognition (CVPR) 2010, San Francisco, CA

- Lin Chen, Qiang Zhang and Baoxin Li, Predicting Multiple Attributes via Relative Multi-task Learning, IEEE Computer Vision and Pattern Recognition (CVPR) 2014, Columbus, OH

- Qiang Zhang, Baoxin Li, "Max Margin Multi-Attribute Learning with Low Rank Constraint," Image Processing, IEEE Transactions on, [accepted]

- Yilin Wang, Qiang Zhang and Baoxin Li, Semantic Saliency Weighted SSIM for Video Quality Assessment, VPQM 2014, Chandler, AZ

- Qiang Zhang, Chang Yuan, Xinyu Xu, Peter Van Beek, Hae jong Seo, and Baoxin Li. Efficient defect detection with sign information of Walsh Hadamard transform. IS&T/SPIE Image Processing: Machine Vision Applications VI. San Francisco, CA, 2013

- Jin Zhou, Qiang Zhang, Baoxin Li and Ananya Das, Synthesis of Stereoscopic Views from Monocular Endoscopic Videos, IEEE Computer Vision and Pattern Recognition (CVPR) 2010 workshop on Mathematical Methods in Biomedical Image Analysis, San Francisco, CA

- Qiang Zhang and Pengfei Xu and Wen Li and Zhongke Wu and Mingquan Zhou, Efficient Edge Matching Using Improved Hierarchical Chamfer Matching, Aug, IEEE International Symposium on Circuits and Systems (ISCAS) 2009, Taipei, Taiwan

- Qiang Zhang and Hua Li and Yan Zhao and Xinlu Liu, Exploration of Event-Evoked Oscillaotry Activities during a Cognitive Task, The 4th International Conference on Natural Computation and The 5th International Conference on Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) 2008, Jinan , China

D.2 MANUSCRIPT UNDER REVIEW

- Qiang Zhang, Baoxin Li, "Relative Hidden Markov Models for Video-based Evaluation of Motion Skills in Surgical Training," Pattern Analysis and Machine Intelligence, IEEE Transactions on