THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

Video-based Pattern Recognition by Spatio-temporal Modeling via Multi-modality Co-learning

By

HAOMIAN ZHENG

A Thesis Submitted in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

July 2012

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

___________________________________(Signature)

_____________________________(Name of Student)

ABSTRACT

The rapid growth of online video content makes it a challenging task to analyze, understand and process video content in real time. Video pattern recognition is emerging as an important research topic in computer vision and communication. Real-time applications such as Internet video searching and video surveillance are popular nowadays. Therefore effective and fast processing approaches are highly demanded. Although the traditional pattern recognition techniques can solve problems for text and image with satisfactory performance, they are subject to certain limitations when processing video due to the large amount of data and time complexity. On the other hand, some statistic models have been proposed for some special video processing applications, however, they cannot handle the general video-based pattern recognition problem. In this thesis, we tackle these problems by addressing three key issues: feature extraction/video representation, indexing, and similarity measurement for classification. The feasibility of the proposed approaches is demonstrated through the experiments on audio-visual speaker identification, video action recognition and gesture recognition.

Firstly we investigate the problem for video feature extraction and representation. Trajectories in high dimensional space are used to represent the video clip and global statistical features are extracted from the trajectory for classification. Based on such feature extraction, we propose two new approaches, Differential Luminance Field Trajectory (DLFT) and Luminance Aligned Projection Distance (LAPD) for the recognition task. For DLFT, we extract the differential signals as features, and then classify the action by supervised learning. For the LAPD approach, we define a new similarity measurement and compute a distance metric to describe the similarity between videos for classification. A potential fusion of the two methods yields more promising properties. Experimental results demonstrate the methods work effectively and efficiently.

Secondly we extend our work by utilizing local spatio-temporal features via indexing. Local features generally contain more statistical information for discrimination. We deal with the spatio-temporal modeling by partitioning appearance space. The proposed approach can capture the discriminative information among different action classes. For trajectory matching solution, we develop a query-driven dynamic appearance modeling method and use localized subspaces to obtain more reliable distance for discrimination. Flexibility is also guaranteed by introducing a warping scheme. The processing is implemented based on an indexing scheme, which is very fast in computation. Simulation results demonstrate the effectiveness of the solution.

Thirdly we focus on improving the pattern recognition performance by proposing novel learning methods. Consider the various features used for video representation, we target on utilizing multiple set of features to jointly solve the recognition problem. We propose a multi-modality distance metric co-learning method. Two set of different features are jointly utilized to generate a better description the video clips. In this way the similarity between video clips is better evaluated and the recognition accuracy is improved. The effectiveness of proposed method is proved by audio-visual speaker identification. Furthermore, to demonstrate the robustness, the proposed method is also applied on digit recognition and text classification. Experiment results show the proposed multi-modality result is better than single modality, together with other previous method in recognition accuracy.

Keywords: Video pattern recognition, video representation, spatio-temporal modeling, multi-modality distance metric co-learning, human action recognition, appearance space indexing.

PUBLICATIONS

Publication List

1. **Haomian Zheng**, Zhu Li, Yun Fu, Aggelos K. Katsaggelos and Jane You, "Video Activity Recognition by Luminance Differential Trajectory and Aligned Projection Distance", accepted by *Handbook on Statistics*, 2012.

2. **Haomian Zheng**, Zhu Li, Yin Yuan, Aggelos K. Katsaggelos and Jane You, "Video-based Human Action Recognition by Dynamic Wrapped Local Spatio-temporal Indexing", to be submitted to *IEEE Transactions on Information and Forensics and Security (TIFS)*, 2012.

3. **Haomian Zheng**, Yan Liu, Timothy T. C. Wong, Qixin Wang, "A SIFT Based Fingerprint for Soft Real-time Visual Localization on Mobile Devices", Submitted to *ACM Multimedia*, 2012.

4. Yin Yuan, **Haomian Zheng**, Zhu Li, "Enhanced Utility Coordination for Video Communication over Ad Hoc Wireless Networks", Submitted to *ACM Multimedia*, 2012.

5. **Haomian Zheng**, Zhu Li, Aggelos K. Katsaggelos and Jane You, "Indexed Spatio-Temporal Appearance Models for Query-driven Video Action Recognition", Accepted in *IEEE International Conference on Multimedia and Expo (ICME)*, 2011.

6. Yin Yuan, **Haomian Zheng**, Zhu Li, Jianwei Huang and Jiannong Cao, "Utility-driven Distributed Transmission Coordination for Video Communications over Ad-hoc Wireless Networks", *IEEE International Conference on Multimedia and Expo (ICME)*, 2011.

7. **Haomian Zheng**, Meng Wang, Zhu Li, “Audio-visual Speaker Identification with Multi-view Distance Metric Learning”, *IEEE International Conference on Image Processing (ICIP)*, 2010.

8. Yin Yuan, **Haomian Zheng**, Zhu Li and David Zhang, “Hand Gesture Recognition by Appearance Space Spline Approximation and spatio-temporal Graph Embedding”, *IEEE International conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

9. **Haomian Zheng**, Zhu Li, Yun Fu, “Human Action Recognition with Luminance Field Trajectory Projection and Alignment”, *IEEE International Conference on Multimedia and Expo (ICME)*, 2009.

ACKNOWLEDGEMENTS

I want to express my sincere thanks to my supervisor, Prof. Zhu Li and Prof. Jane Jia You, whose expertise, understanding, and patience, added considerably to my graduate experience. I appreciate their vast knowledge and skill in many areas and their professional supervision. It is my great pleasure to become the first research student of Prof. Li, and I want to thank him for supporting me over the years, both mentally and academically. I can hardly imagine the current work output without his help and support. Prof. Jane You gave her kind guidance suggestions on guiding my research work, and took care of my study after Prof. Li's leaving. Her experience and research spirit impressed me deeply and this will be definitely helpful in my future work.

I also want to give my gratitude to Prof. George Baciu, Prof. Pong Chi Yuen and Prof. Xiaoyi Jiang to be my Board of Examiners, for their time and involvement of my PhD examination.

I would also express my gratitude to the other members of Dr. Li's research group, Dr. Wen Ji, Mr. Bo Liu, Mr. Hao Xu, Mr. Xinchao Wang, Ms. Zhi Ye and Mr. Xin Chen, for the assistance they provided during my Ph.D. study. Also, I would like to thank Prof. Yan Liu's research group, Mr. Yang Liu, Ms. Shenghua Zhong, Mr. Yao Zhang. I spent my final year in this group and had a happy time. I also would like to thank all my teachers from whom I learned so much in my long journey of formal education.

Specially thanks go to Prof. David Zhang, Prof. Jiannong Cao, Prof. Qixin Wang, Prof. Dan Wang, Prof. Ajay Kumar and Dr. King Hong Cheung at the Hong Kong Polytechnic University.

Furthermore, I acknowledge my gratitude to Dr. Qijun Zhao, Dr. Dennis Guo, Dr.

Lin Zhang, Dr. Dongmin Guo, Dr. Bo Peng and Dr. Xiaofeng Qu from Biometric Research Center, Dr. Zhiwei Qin, Dr. Duo Liu, Dr. Yi Wang, Mr. Zhizhao Feng Mr. Yufei Wang and so on from Department of Computing and Dr. Xuran Li, Dr. Yang Wu from Department of Electric Engineering, who shared with me the pleasure of the Ph.D. study at the Hong Kong Polytechnic University.

I must acknowledge Prof. Aggelos K. Katsaggelos, for offering me the opportunity to visit IVPL in EECS Department of Northwestern University. His truly scientist intuition and invaluable guidance inspires and enriches my intellectual maturity that I will benefit from, for a long time to come. I would also thank the colleagues there, Dr. Derin, Babacan, Dr. Fan Jiang, Dr. Xin Xin, Dr. Lei Song Ms. Jingnan Wang, Dr. Eren Soyak, Dr. Martin Lusse and Dr. Lu Fang, for their kind suggestions and ideas in my research.

I offer my regards and blessings to Dr. Xian-sheng Hua and Dr. Meng Wang, for inviting me to visit Microsoft Research Asia in 2009. Thanks for the assistance from Media Computing Group members, Dr. Linjun Yang, Dr. Tao Mei, Dr. Jingdong Wang, Dr. Lifeng Shang, Dr. Kuiyuan Yang, Dr. Xiangang Cheng, Mr. Jin Sun, and all of those who supported me in different respects during my visit at MSRA.

I also recognize that this thesis would not be completed without the financial assistance from the Hong Kong Polytechnic University. I appreciate Prof. Li and the Department of Computing for offering me the travel grants to attend several international conferences. I acknowledge the grant for Research Student Attachment Program from the Hong Kong Polytechnic University for giving me the financial support to visit Northwestern University.

Finally, I want to thank my wife, Ms. Yin Yuan, for her continuous care and support. I also give my thanks to my parents. They educated and guided me and have watched over me every step of way. I want to thank them for their endless love, support, and encouragement through my entire life, for letting me pursue my dream for such a long time, and far away from home, and for giving me the motivation to finish this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

The explosive growth of video content, especially those in online video repositories like Youtube and those recorded from millions of surveillance cameras, are presenting a challenge to real-world video analysis and understanding systems. An urgent need is to develop effective and efficient video understanding solutions for a variety of pattern recognition applications, such as in event detection, action recognition, person identification, which requires both robustness in recognition and efficiency in computation and possibly real-time performance.

Video content analysis is the capability of automatically analyzing video to detect and determine temporal events not based on a single image but a video sequence. This technical capability is used in a wide range of domains including entertainment, health-care, retail, automotive, transportation, home automation, safety and security [11] [83]. The algorithms can be implemented as software on general purpose machines, or as hardware in specialized video processing units.

Video patterns are high-level semantic concepts that humans perceive when observing a video sequence. Video content and event understanding attempts to offer solutions to the problem of detecting the human perception of content with a computer perception. The major challenge for content analysis and event understanding is how to effectively translate low-level input into a semantically meaningful event description [47].

Video pattern recognition is a kind of high level task in computer vision. It relies on sufficient solutions to many lower level tasks such as denoising, edge detection, optical flow estimation, object recognition and tracking. The maturity of many solutions to these low-level problems has spurred additional interest in utilizing them for higher level tasks such as video event understanding.

Another reason for the large amount of interest in video-based pattern recognition is the promise of intelligent systems outfitted with inexpensive cameras enabling such applications as active intelligent surveillance, summarization and indexing of video data, and human computer interaction. There are various applications around such as healthy care and kinetic analysis.

The problem of video pattern recognition is still challenging due to several reasons. The noise brought by different scale, rotation and illumination will confuse the computer, with uncertainty and large appearance variance in the particular events. On the other hand, similarity in the appearance of different events also degrade the performance. Therefore, it is a critical problem in video pattern recognition on how to effectively separate the discriminative information together with remove the noises which will result in misunderstanding. For simplicity, we define two basic questions for video pattern recognition, how to efficiently extract and preserve the discriminative features, and how to effectively classify the extracted features into the correct categories. We define the first class of question as video representation, and the second question as modeling.

In this thesis, we address the challenges of video-based pattern recognition from three aspects, video representation, spatio-temporal modeling and machine learning algorithms. In video representation we propose trajectory representation in our solutions, and present different methods to preserving information when reducing the number of dimensionality of data. Spatio-temporal modeling is highly emphasized in this work, appearance and temporal modeling are respectively considered in the solution to achieve better classification result. For machine learning problems, we propose multi-modality distance metric co-learning and improve the recognition accuracy. Basically we are confronted with the following problems in the processing.

Firstly, the video sequence is usually composed of hundreds of frames, each of them is composed by millions of pixels. The data amount is too large to efficient content understanding. It is well-known that lots of redundancy exist in the video sequence. Therefore in the first step we focused on extracting representative features from the video sequence. Down-sample and vectorization are applied to reshape each frame, and a trajectory is constructed by connecting the image sequence. The amount of data is dramatically reduced in this step but still more than necessary.

Secondly, there is no obvious mark to show the location of the discriminative information. In this thesis we propose many different spatio-temporal modeling schemes to address this problem. Features in temporal domain is highlighted in the proposed Differential Luminance Field Trajectory (DLFT) approach. Appearance modeling is constructed in the proposed Luminance Aligned Projection Distance (LAPD) method. Localized appearance indexing and warping-based solution are proposed to further improve the recognition accuracy.

Finally, we also consider on improving the recognition performance by exploring new machine learning techniques. Instead of considering only one set of features, multi-modality case is also investigated in this thesis. Distance metric co-learning is demonstrated to be contributive in various applications.

The rest of this chapter is organized as follows, an overview of video-based pattern recognition is presented in Section 1.1. Section 1.2 presents the unified research framework. Section 1.3 summarizes the contributions of this thesis. Finally we give an outline of the thesis in Section 1.4.

1.1 An Overview of Video-based Pattern Recognition

In this section, we present some related techniques in video pattern recognition literature. They are basically divided into three categories, video representation, spatio-temporal modeling approaches and machine learning methods.

Video representation refers to translating video sequences into intermediate units understandable by spatio-temporal models. In section 1.1.1 we provide some representation approaches.

Spatio-temporal modeling is a critical section in the video pattern recognition problem. Given input from the representation layer, the model should categorize the video sequence into several pre-defined classes. Spatio-temporal modeling has already received a lot of attention in the computer vision research community, and we will briefly introduce these methods with our comprehensive understanding and analysis in section 1.1.2.

Machine learning is a classic research topic, and also the final step in video-based pattern recognition. Every solution should have a section on the machine learning and we will present some popular techniques in section 1.1.3.

1.1.1 Video Representation

Representation is the organization of low-level inputs into various primitives representing the abstract content of the video data. It is motivated by providing an intermediate summarization of the video content. Compared with modeling issues, representation is not highlighted in the literature. However, every research work should consider how to present the low-level features in an efficient way. This decision is the output of representation phase and is an integral part of the video pattern recognition processing.

Researchers were interested in pixel-based representation in the past a few years. Pixel-based representations utilizes abstraction schemes that rely on single or group of pixel features such as texture and color moment. Motion history image [104] and gradient histogram [9] are examples of pixel-based representation.

Intuitively video content can also be composed by a group of different objects. Therefore object-based representation is becoming an alternative solution. Low-level input is abstracted and object properties, such as speed, position and trajectory are used as representation [25] [35] [63] [82]. Silhouettes are another popular object-based representation, which is widely used for action recognition [6] [72].

Another group of representation can be categorized into concept-based. The idea is that the daily video content is not composed by pixels, and can hardly described by a group of object. Instead, it should be described by some semantic concept. Scale Invariant Feature Transform (SIFT) is firstly proposed in [53] and used as a definition of "word" as a representation. It is widely applied in image processing and Spatio-temporal Interest Point (STIP) is developed in [44] for video representation.

1.1.2 Spatio-temporal Modeling

As presented in the previous section, spatio-temporal modeling is the complementary problem to representation. The modeling phase targets on seeking formal ways to describe and recognize specific video content in a particular domain given the choice of a representation scheme. A particular spatio-temporal model is chosen based on both the capacity for representation in a particular domain and the capacity for recognition of these content as they appear in the video sequence input.

Spatio-temporal modeling methods can be categorized into many different ways. Most of the research works propose novel modeling schemes to improve the performance. The model can be either deterministic or probabilistic, either generative or discriminative. Depend on different applications, the models also vary a lot.

However, such kind of division did not fully capture the diversity of event modeling approaches in the video pattern recognition literature. Therefore we further categories the models into "state modal" and "semantic models". Noted that such kind of category is not meaning that ever model should be exclusively include into one class.

We defined the first class of these models as "state models" for the reason that they concentrate on specifying the state space of the model. Often, this state space is reduced or factorized using semantic knowledge. This class of approaches includes finite-state machines (FSMs) and the set of probabilistic graphical model approaches. The existence (under some structural assumptions) of efficient algorithms for the learning of parameters from training as well as recognition motivates the choice of these models to model video patterns.

Higher level semantics include ordering information (including partial ordering), and complex temporal, spatial, and logical relations. These properties become important when the event domain includes high-level events, which are best expressed in qualitative terms and natural language. To this end, a group of modeling formalisms that we defined as “semantic models” have been proposed, which enable explicit specification of these complex semantic properties. Among these are Petri nets (PNs) and grammar models as well as constraint satisfaction and logic-based approaches. These models are usually fully specified using domain knowledge and are not usually learned from training data.

An effective model will be definitely helpful to the pattern recognition. Besides such models, machine learning methods will also affect the performance. We will present a review of such methods in 1.1.3.

1.1.3 Similarity Measurement and Classification

Generally, the video pattern recognition cannot be solved by classic pattern recognition method directly. The machine learning methods could be contributive in both representation and modeling phase. Traditional approaches such as support vector machines, nearest neighbor classifiers and neural networks, are widely applied in the processing. There is no direct connection between such classifier and semantic knowledge, so they are generally “blind” in the classification procedure.

The class of techniques in this section is not quite spatio-temporal models, in the sense that they do not consider the problem of video content representation. Instead they focus on the pattern recognition problem. The main advantage of the classifiers in this category is that they are well understood. Usually, they may be fully specified from a set of training data. These approaches are usually simple and straightforward to implement. This simplicity is afforded by excluding semantics (i.e., high-level knowledge about the event domain) entirely from the specification of the classifier.

There are many examples of pattern-recognition methods for event recognition in the literature. Nearest neighbor based classifiers are widely used in [6] [9] [73] [104] . Support

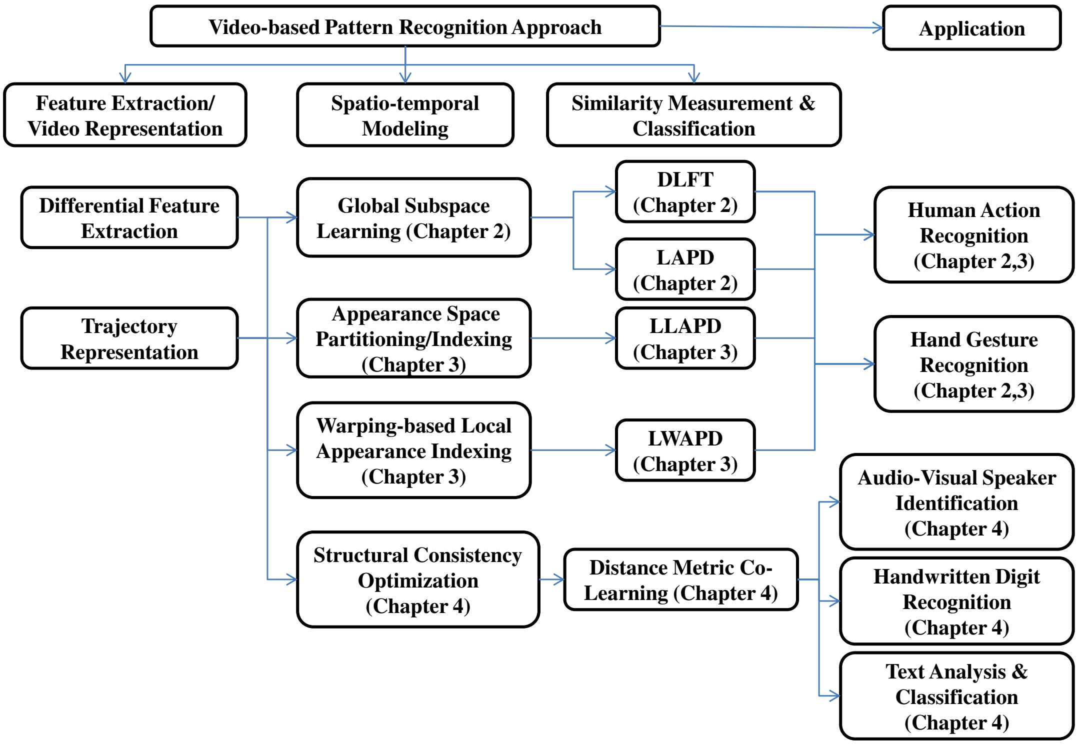

Video-based Pattern Recognition Approach
Application
Feature Extraction/ Video Representation
Spatio-temporal Modeling
Similarity Measurement & Classification
Differential Feature Extraction
Trajectory Representation
Global Subspace Learning (Chapter 2)
Appearance Space Partitioning/Indexing (Chapter 3)
Warping-based Local Appearance Indexing (Chapter 3)
Structural Consistency Optimization (Chapter 4)
DLFT (Chapter 2)
LAPD (Chapter 2)
LLAPD (Chapter 3)
LWAPD (Chapter 3)
Distance Metric Co-Learning (Chapter 4)
Human Action Recognition (Chapter 2,3)
Hand Gesture Recognition (Chapter 2,3)
Audio-Visual Speaker Identification (Chapter 4)
Handwritten Digit Recognition (Chapter 4)
Text Analysis & Classification (Chapter 4)

Figure 1.1. The Unified Research Framework.

vector machine (SVM) is applied in [16] [24] [64] [66] [93]. Boosting based method is tested in [15] [45] [55] [60] [67] [77].

1.2 The Proposed Unified Framework

In this section, we present the unified research framework for the proposed techniques. Figure 1.1 illustrates the sketch of our research framework.

In this thesis we target on solving the video-based pattern recognition problem. The solution is composed by three different aspects, feature extraction/video representation, spatio-temporal modeling, similarity measurement and classification.

For video representation scheme, we convert each of the video sequence into a trajectory in a very high dimensional space. The problem for comparison between videos will be translated to trajectory matching.

Given the trajectories, spatio-temporal modeling is highly emphasized in this thesis. We propose different approaches to solve the modeling and matching problem. The proposed methods are tested by human action recognition and gesture recognition respectively.

For the machine learning block, we investigated the statistic feature in the video sequences and focused on finding a solution to fully utilize these features. We present our proposed multi-modality distance metric co-learning method and demonstrated it with three different applications, audio-visual speaker identification, handwritten digit recognition and text classification.

1.3 Contributions and Related Publications

The contributions of this thesis can be divided into three different aspects on solving the video pattern recognition problem, trajectory-based representation, novel spatio-temporal modeling, and improved machine learning matching solution. The highlights are summarized as follows.

- For video representation, we propose a trajectory base representation scheme. The video similarity is then converted to an evaluation equivalent to trajectory distance. The processing is simple and fast, which can achieve a balance between performance and processing time. In our related publications, such as [101], [106], we are using such representation methods.

- Spatio-temporal modeling for video pattern recognition is the main contribution of this thesis. We investigate both global and local modeling scheme in this work. For global consideration, we propose Differential Luminance Field Trajectory (DLFT), which takes advantage of the differential signal as discriminative features. The Luminance Aligned Projection Distance (LAPD) is proposed to evaluate the similarity between video sequence by finding the optimal match of the trajectories. The DLFT and LAPD work is published in [105] and [108] respectively.

To further improve the performance, we propose local indexing scheme by partitioning the original appearance space into subblock and apply subspace learning individually. The local statistical information is proved to be contributive for classification. Local LAPD and warping LAPD is proposed as two solutions, with better recognition accuracy than global solutions. The local LAPD is included in *International Conference on Multimedia and Expo* (ICME). Details can be found in [106]. The warping LAPD is presented in [109].

- We propose a novel multi-modality learning method to solve pattern recognition problem. It is a general algorithm which can be applied on text, image, video and other kind of data. Motivated by the benefits brought by multiple group of features, we present an algorithm on utilizing the mutual information from different set. Cross validation is proved to be helpful on classification, multi-modality solution also offers better solutions than single-modality one. This methodology is published in *International Conference on Image Processing* [107].

1.4 Thesis Organization

The rest of this thesis is organized as follows.

- In Chapter 2, we propose a trajectory-based feature extraction and video representation scheme, followed by spatio-temporal modeling for video action recognition. We present two approaches focused on temporal feature and spatial feature respectively. Maximum Likelihood is used for decision. The numerical results from different dataset are competitive with the ones in the literature, so that the robustness of the proposed method is also guaranteed.
- In Chapter 3, we present our proposed spatio-temporal modeling approach. A local indexing scheme, which can improve the previous solution by utilizing local statistical information, is proposed to improve the recognition accuracy. Fast space partitioning is applied to achieve a balance between recognition accuracy and processing

time. Warping-based solution is also proposed to provide more flexibility on trajectory matching. Experiment result shows that the performance is successfully improved by the proposed approaches.

- In Chapter 4, we solve the video-based pattern recognition problem by improving the pattern recognition solution. Multi-modality distance metric co-learning is proposed by utilizing structural consistency between different set of features. The proposed solution is demonstrated by experiment to be effective in different applications.
- In Chapter 5, we present conclusions and propose several potential future directions of research arising from this work.

CHAPTER 2

DIFFERENTIAL LUMINANCE FIELD TRAJECTORY AND ALIGNED PROJECTION DISTANCE: SOLUTIONS FOR VIDEO-BASED HUMAN ACTION RECOGNITION

2.1 Overview

With the development of computing and communication technologies, video content analysis is becoming one of the most popular research areas in computer vision and machine learning. Video action pattern recognition has wide applicability in video surveillance, s-ports, entertainment, searching, human-computer interaction and many other activities in daily life. Basically, the problem can be defined as determining a query action into several pre-defined ones. The set of actions contains a semantic meaning in our daily life, such as running, clapping, or jumping.

Actions can be categorized into different spatio-temporal patterns. Therefore the extraction of appropriate features is critical to solve the problem. Various kinds of features such as luminance points [33], human body detection [98], spatio-temporal interest points [44] were proposed in the literature and proved to have good performance on discriminative action recognition.

Although the central problem looks very simple, there are several quite challenging sub-problems, which are the subjects of intense research. Generally, these sub-problems can be categorized into low-level pre-possessing, human body representation, and subspace learning. The low-level pre-processing applied on the original video clip leads to the information extraction which is used as a representation of the human body for action recognition.

The subspace learning basically aims to finding a subspace projected to which the discrimination can be preserved while reducing the dimensionality. A subspace can be learned in various ways to train a system, which can recognize the query video clip automatically [49] [46].

Some people-computer interactive systems are also developed in this task. In such a kind of a system, people can define some special points to help recognition. For example, in [98], the hands, feet and head of people are manually marked by the user before training and recognition. With this prior information, the performance is usually better than that of the automatic systems.

In this chapter we propose two different methods, both of which are based on luminance spatio-temporal features, to recognize different actions automatically. The first approach is the Differential Luminance Field Trajectory (DLFT). Luminance video frames are vectorized and projected into a high dimensional space thus forming a trajectory which is generated to represent the video. Differential values of the luminance trajectory are used for real-time on-line classification. In our second approach, Luminance Aligned Projection Distance (LAPD), a new measurement for similarity is proposed by introducing an appropriate subspace and using an optimal alignment. A distance metric is constructed in this way and a KNN classifier can easily be applied for recognition.

The chapter is organized into the following sections. In section 2.2, we briefly review some of the popular techniques in the human action recognition literature. The formulation for our proposed methods is given in section 2.3. Solutions for proposed methods is presented in section 2.4, simulation results are shown in section 2.4 and compared with other recent works. Finally conclusions are drawn in section 2.6 and further works are included as well.

2.2 Related Works

As an important area in computer vision, the human action and activity recognition have received much attention in recent years. A comprehensive review of this research topic has been presented in a number of survey papers, e.g., [1] [20] [83]. In this section, we mainly focus on discussing the most critical processing in this special problem.

Pixel values can be directly obtained from an image or a video clip. So the optical flow representation, which is based on the moving pixels, has been widely used as a simple representation of the video by a lot of researchers [23] [33]. In this approach, the idea is to directly use the optical flow to derive a video representation which can be used for recognition. So motion detection and analysis from video compression work have also been combined into this technique. For example, Motion Energy Image (MEI) [8] and Motion History Image (MHI) [9] have been proposed to describe the motion information.

In general, a class of approaches for human action recognition analysis are based on the modeling of the extracted features from the video sequences. The modeling and learning of the extracted features is the critical part of the problem, in improving the accuracy of the recognition. Some popular techniques include optical motion detection, 3-D volume representation, temporal modeling, Hidden Markov Model (HMM) training, Dynamic Time Warping (DTW) and multi-view subspace learning. We offer a brief review of these techniques in the following several paragraphs.

The appearance based feature representation is not robust with respect to background changes such as scaling and rotation. Also, the failure on handling occlusions and cloth changing limited the application on these methods. Space-time interest points and their trajectories for action and activity analysis are quite popular in the recent literature [46] [62]. The main strength of this representation is the robustness to occlusions, since there is no need to detect or track the human body or hand. A dictionary can be constructed by a bag-of-words approach and therefore the image or video can be represented as statistical information of words.

Temporal properties have been proved to contribute a lot towards action classification. Compared with traditional 3-D modeling, a 4-D (x, y, z, t) action feature model (AFM) was proposed for representation and recognition of action from arbitrary views [96]. Temporal features are also highly emphasized in [86] for creating intelligent robot systems. By utilizing Conditional Random Fields (CRF) and applying discrimination training, the algorithm is proved to be effective.

Researchers also applied Hidden Markov Models and their variants for better analysis of their temporal behavior [79]. The general methodology was to learn the appearance model of the human body or hand and match it explicitly to images in a target video sequence for action and gesture recognition [95]. This approach is highly dependent on the features extracted from the video. Different representations also has different models. In [28], actions in video clip are treated as 3-D shapes induced by silhouettes in the space-time volume and properties of the solution for Poisson equation was utilized to extract the features such as action dynamics, shape structure and orientation. The method is proved to be fast and robust to partial occlusions and can be applied to low-quality videos. Similarly, in [88] an exemplar-based Hidden Markov model (HMM) was proposed and this model took advantage of dependencies between three dimensional exemplars. Furthermore, a template-based method, named the Maximum Average Correlation Height (MACH), was proposed in [68]. By capturing the intra-class variability, the single action class is simply and carefully modeled after analyzing the response of the MACH filter.

Dynamic Time Warping is a traditional approach in speech recognition, which is used to align audio sequence of different duration. This technique was recently applied for human action recognition as well [19] [65]. This method is proved to be robust to scaling and rotation, and also has good performance given only very low-dimensional features, but its application is limited to only a few action patterns.

Instead of building models for only one set of the features, there are some approaches that focused on both temporal and spatial domains. A more comprehensive understanding can be obtained during such a process. In [39], a spatio-temporal volume modeling based solution is investigated and proved to be insensitive to image formation variations. In [90], a new approach, which is composed of a 2-layer statistical field model, was proposed and demonstrated to be robust to occlusions. Besides robustness, the structure was also more flexible to image observations, which made the method robust to clutter as well. In [43]

Canonical Correlation Analysis (CCA) is used to measure the similarity of any two image sets for robust object recognition. Correlation information is also considered to be helpful for recognition. After this, in [42], a method is also applied for hand gesture recognition by combining feature selection and the Tensor Canonical Correlation Analysis (TCCA) learning process. Tensor work has also been applied for gait recognition in [80], combined with Gabor features contained in the gait sequence.

2.3 Problem Formulation

In this section we describe the problem formulation, along with several pre-processing steps: video representation, noise reduction and subspace learning. Challenges in the human action recognition tasks include human detection and representation, motion understanding and analysis. By solving these problems with appropriate algorithms, the signals can be prepared for learning and recognition.

Video clips are composed of frames which consist of by pixel values. It is a challenging task to detect the human body in video sequences, especially with large visual variations and occlusions. Originally, researchers treated human as a single object in the frame so that the human body can be separated from the background. A number of solutions based on this idea have already been proposed in the literature. Traditional methods focused on detecting and recognizing different human actions, such as in [18] and [103]. The main techniques involved are the so-called "object-extraction-based" method [76], which extracted a certain object by image processing techniques, such as edge detection and object segmentation processing. However, the appearance of human body in video sequences may not be very concrete and is easily corrupted by noise. This approach suffers from lack of robustness to lighting, nature of the background, and occlusions.

To make algorithms more robust, different kinds of video representations were introduced to capture the invariance in the video, such as local image features or spatio-temporal interest points in [44], which provided a compact and abstract representation for patterns within a given image. The applications included object detection, tracking, and segmentation. The performance was demonstrated to be robust for variations of background. The so-called Scale Invariant Feature Transform (SIFT) points were then proposed in [53], and a method was designed for extracting distinctive invariant features based on the SIFT points. The SIFT points were selected by calculating the Difference of Gaussians at every pixel and representing as descriptors in different directions. Points of interest can also be encoded as a histogram [72] and this kind of representation is combined with a Support Vector Machine (SVM) [12] or some other probabilistic model. Using similar ideas, a generative graphical model in [58] used the interest points for human action recognition. This method analyzed the human action directly in the space-time volume without explicit motion estimation [73].

On the action understanding side, people focused on detecting the type of action by motion analysis. After extracting the human in the video clip, the human can be represented by several special parts, such as, arms and legs. The action is analyzed by detecting the motion of these parts and models for different actions can be learned from the given motions during the training process [98]. With the various backgrounds and different viewing angles, how to effectively detect the critical points on human body becomes the main difficulty for these approaches.

To avoid the detecting difficulties in critical points, appearance based approach is proposed to solve more general problems. Subspaces can be learned from the training video clips and used to model the query ones. Traditional subspace modeling includes Principle Component Analysis (PCA) [84] and Linear Discrimination Analysis (LDA) [4] and so on. Linear Projection Preserving (LPP) is proposed in [31] to build a graph on understanding the neighboring information.

Another popular approach is to treat the human action as a sequence and learn the model from the difference in the temporal domain. Besides, in [30], a non-linear principal curve approximation was developed. Intuitively, it is a curve passing through the center of the data points cloud, with a smoothness constraint. In [40] and [41], it was demonstrated that as long as the second moment of the data points cloud is finite, there must be a principal curve, and an iterative polygonal principal curve learning algorithm was developed.

In this work, we model human action video clips as manifolds in the scaled appearance space over time. Video clips of different human actions performed by different subjects under different image formation conditions span a space with complex structure and relationships. By scaling the original video frames into icon images, the local noise can be effectively attenuated while the information about the action is maintained. The formulation can be divided into video representation, subspace learning and matching problem.

2.3.1 Video Representation

Video representation is the first and also one of the most important sub-problems in video pre-processing. A good representation should include the key point and useful information for discrimination while discarding unnecessary information.

Generally, in video processing, video frames are usually represented as a matrix. In our method, we use the luminance information to keep the data in every single frame, with a vector structure. To simplify pre-processing, pixel values are directly extracted as features. The video frame is first down-sampled to a smaller icon to reduce spatial redundancy together with noises, and then the icon is projected into a high dimensional space and become a point. In this way, different video clips of different human actions performed by different subjects under different image formation conditions span a space with complex structure and relationships. The spatial features in the clips are kept in a vector form while the temporal ones are included in the trajectory as well.

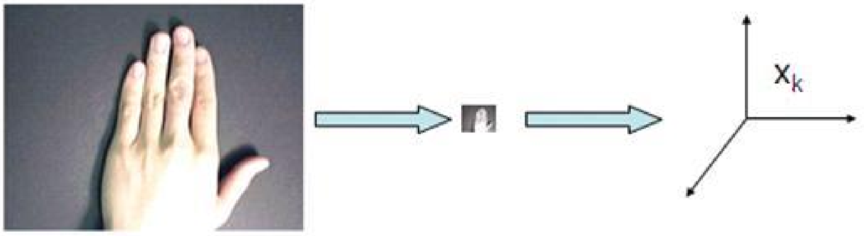


Figure 2.1. Video frame representation by down-sampling and projection

Considering a video clip which contains n frames, with $W \times H$ pixels in each frame, the kth frame F_k can be represented as a point in the space $\Re^{W \times H}$. Actually the frame of size $W \times H$ still contains more information than necessary, so down-sampling it will reduce the number of elements while keeping adequate information for recognition. The down-sampling step reduces the original frame down to a smaller $w \times h$ one. By down-sampling each frame can be further represented as a point in a space of smaller dimension, i.e., $\Re^{w \times h}$. After this processing, the trajectory still contains sufficient statistical discriminative information for classification. The left part of Fig. 2.1 shows the down sample processing.

2.3.2 Dimension Reduction

To further simplify the processing, another pre-processing step is introduced by subspace learning. In this step a global subspace is learned. By projecting the every sample points to the subspace, the discriminative information in the set is maintained while the number of dimension is reduced for faster processing. A global PCA [84] is applied here to reduce the dimensionality of the space, consider a n-frame video sequence, given a frame $F_k \in \Re^{w \times h}, k \in [1, n]$, the subspace learning can be expressed as:

$$x_k = AF_k = [a_1, a_2, ..., a_{w \times h}]F_k, a_j \in \Re^d \tag{2.1}$$

where the subspace projection A, of size d by $w \times h$, is obtained from an unsupervised local learning, with the objective of preserving the maximum amount of information, while keeping the number of dimension an acceptable level. The global subspace projection is shown in the right part of Fig. 2.1. Each a_j in Eq. 2.1 is a $d \times 1$ column vector of matrix A.

Figure 2.2 shows 3 groups of curves, each for a different human action in the Cambridge hand gesture dataset in 3-D space. In the figure, video clips containing different actions have different shapes, and as one can judge some actions are clearly different, while for others it is rather difficult to distinguish them since the 3-D view cannot offer enough visual information for doing so. Actually the geometry information of these curves already

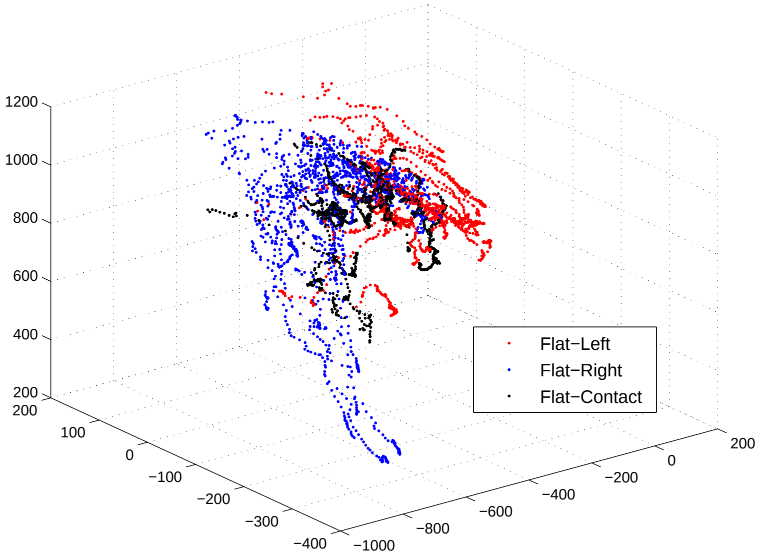


Figure 2.2. 3-D example for Cambridge Handgesture Dataset

contains sufficient statistics to recognize the different human actions. In the next 3 sections, we will propose our approach based on such kind of statistical information.

2.3.3 Maximum Likelihood Detection

Video clips are represented as trajectories in a high dimensional space R^d after the preprocessing and dimension reduction. The representation is still not simple enough to discriminate the different action classes.

Assume that the training point set $\{x_1, x_2, ..., x_n\}$ in the R^d space is with Gaussian distribution, then we can obtain the mean m_x and the variance σ_x respectively. Given a query frame q, and a training frame x, the likelihood that q and x have same action label is also under a Gaussian distribution, i.e.,

$$\mathcal{L}(q; x) \sim N(m_x, \sigma_x^2) \tag{2.2}$$

Since x is a Gaussian Mixture, the likelihood can be further computed as

$$\mathcal{L}(q;x) = \mathcal{L}(q;m_x,\sigma_x) = \frac{1}{(2\pi\sigma_x^2)^{d/2}} e^{-\frac{1}{2}(q-m_x)^T\sigma_x\sigma_x^T(q-m_x)} \tag{2.3}$$

In this way, given a query trajectory $q(t)$ and a training sequence $x(t)$, $t = 1, 2, ..., n$, the likelihood between q and x can be computed as:

$$\mathcal{L}(q(t);x(t)) = \prod_{t=1}^{n} \mathcal{L}(q(t);x(t)) \tag{2.4}$$

If there are totally k training trajectories, we can compute the k likelihood and then decide the action label of $q(t)$ by using Maximum Likelihood decision as follow,

$$k^* = \arg\max_k \prod_{t=1}^{n} \mathcal{L}(q(t);x(t)) \tag{2.5}$$

However, in the trajectory matching, another challenge is to match up the sequences with different durations. Trajectories with different length may have different matching options. If we consider every possible matching option, the growth indifference of duration will make the matching complexity grows exponentially, which is computationally prohibitive. In this work we compute point-to-point likelihood, which strictly keep the temporal information in the trajectory. It is constrained that points must be matched according to order. No skipping, repeating or crossing is allowed for matching. In this way, the matching is simplified into a linear level. In order to find the optimal matching, likelihood of all the possible matching offset between two trajectories are computed. The maximal value is selected as the best evaluation. Therefore, the likelihood computation is corrected into,

$$\mathcal{L}(q(t_q);x(t_x)) = \begin{cases} \min\limits_h \prod\limits_{t=1}^{n} \mathcal{L}(q(t_x+h);x(t_x)) & \mathrm{t}_q > t_x \\ \prod\limits_{t=1}^{n} \mathcal{L}(q(t_q);x(t_x)) & \mathrm{t}_q = t_x \\ \min\limits_h \prod\limits_{t=1}^{n} \mathcal{L}(q(t_q);x(t_q+h)) & \mathrm{t}_q < t_x \end{cases} \tag{2.6}$$

In this way, a time align process is performed and the matching computation is simplified to a linear level. To achieve optimal matching, likelihood of all the possible matching offset between two trajectories are computed. The maximal likelihood is selected as the best evaluation. Then the decision can be made as,

$$k^* = \arg\max_k \sup_h \prod_{t=1}^{n} \mathcal{L}(q(t_q); x(t_x)) \tag{2.7}$$

2.3.4 Luminance Aligned Projection Distance Approach

The Bayesian-based likelihood solves the trajectory matching problem effectively, but not efficiently. The computation for likelihood in Eq. 2.3 is very complicated and limits the method to many real-time applications. In this section we propose a simplified version, the Luminance Aligned Projection Distance (LAPD) approach, to efficiently solve the matching problem. In the proposed approach, we defined the similarity between two video trajectories as the trajectory distance. In other words, the smaller the trajectory distance is, the more likely the two trajectories have same action labels. In this way, the effectiveness in computing trajectory distance will result in a satisfactory recognition result.

Given a training set after dimension reduction, the parameter d, m_x and σ_x are constant for every query clip. So the computation of likelihood can be simplified by removing the first multiplier factor in Eq. 2.3, i.e.,

$$\mathcal{L}(q; x) = \mathcal{L}(q; m_x, \sigma_x) = e^{-\frac{1}{2}(q-m_x)^T \sigma_x^{-1} \sigma_x (q-m_x)} \tag{2.8}$$

The exponential function is a monotone increasing function, and the detection of maximum likelihood is equalized to find the minimum Mahalanobis distance from query clip q to m_x. The distance measurement is more reliable when a subject is repeating same action under different illumination conditions or in different background. Especially, the subspace can be directly obtained by decomposing the covariance matrix, $S = \sigma^{-1}\sigma = A^T A$. Based on this observation, we propose a distance-based approach to detect the maximum likelihood. Intuitively, the distance between two trajectories, is believed to be an effective and reliable measurement for similarity. Samples which have similar content should have smaller distance, as compared to those with dissimilar content. To make the decision more promising, we use a KNN classifier rather than maximum likelihood detection, which is equivalent to a 1-NN classifier. The classifier is applied after the distance metric is computed to label the query clip.

2.3.4.1 Distance from Point to Trajectory

In this section the distance between a single point and a trajectory in a d-Dimensional space, with subspace modeling A, is discussed.

Basically, the distance between two points in the subspace A can be defined as Eq. 2.9.

$$d(x, y) = ||A(x - y)||^2 = (x - y)^T A^T A(x - y) \tag{2.9}$$

where both x and y are points in $\Re^d$. Specially, when A is a unit matrix, the $d(x, y)$ is the Euclidean distance. Also, when A is the variance in the Gaussian training set, the distance definition becomes a special variation of the likelihood defined in Eq. 2.3.

Furthermore, consider a point x and a trajectory Y composed of a group of points $\{y_1, y_2, ..., y_n\}$. Then similarly, the distance from a point x to the trajectory can be defined as the minimal point-to-point distance, i.e.,

$$d(x, Y) = \min_i d(x, y_i) = \min_i ||A(x - y_i)||^2 \tag{2.10}$$

2.3.4.2 LAPD: Distance Between Trajectories

The distance between trajectories shows the similarity between two video clips. In previous sections we compute the point-to-point and point-to-trajectory distance. In this section we compute the inter-trajectory distance in a similar way.

To compute an effective distance which gives reliable similarity representation, the Luminance Aligned Projection Distance (LAPD) is proposed based on the following idea: given a pair of trajectories, finding an optimal matching offset in the longer trajectory started at where the afterwards average point-to-point distance is minimized.

Suppose trajectories are denoted as $x_{j,k}(t)$, for curve j belonging to action class k, and for each class, there are $j = 1..n_k$ curves, t is the frame index which varies from 1 to n. Then for an unknown video clip trajectory $y(t)$ with m frames, and a known action video clip $x(t)$ of n frames, assuming $m < n$, the LAPD between x and y is defined in Eq. 2.11.

$$d_{\text{LAPD}}(x, y) = \min_h \frac{1}{m} \sum_{t=0}^{m-1} ||A(x(t+h) - y(t))||^2 \tag{2.11}$$

where the h is denoted as the difference of frames between two trajectories, i.e., how many frames are the two video clip different? In this way we have h possible matching scenarios:

Scenario 1: Match x(1), x(2), x(n) to y(1), y(2), y(n), respectively.
Scenario 2: Match x(1), x(2), x(n) to y(2), y(3), , y(n+1), respectively.
...
Scenario h: Match x(1), x(2), x(n) to y(h+1), y(h+2), , y(h+n), respectively

Let us assume that we have K action classes and each has $j = 1..L$ training clips, $x_j^k(t)$. Then for an query clip $y(t)$, recognition can be implemented based on the minimum LAPD,

$$k^* = \arg\min_k \min_j d_{\text{LAPD}}(y(t), x_j^k(t)) \tag{2.12}$$

For each incoming query video clip C, we calculate the distance between C and those clips in the training set which contains N training samples. An $N \times 1$ distance array D_N is generated with different action labels. After sorting the entries of the distance array D_N, the M smallest values are selected from the training distance array with the corresponding action labels. Therefore, given the first M labels, voting is applied to count the number of labels for each action class. The final decision is based on the label with the most votes.

In this method, we focused on finding the relationship between the two trajectories in subspace. Instead of computing the distance directly, a best matching point is firstly found by trying every possible offset in matching the two trajectories which minimize the distance between them. The spatial information is maintained in the trajectory coordination, while the temporal features are kept by continuous point-to-point matching. This processing removes the effect brought by noise, and is proved to be robust against some other factors such as scaling and background changes.

2.3.5 Differential Luminance Field Trajectory Approach

The computation of distance in previous section spends less effort than likelihood, but still complicated. In this section we provide a faster solution by utilizing the differential features in the trajectory. Consider a query clip with n frames, and there are h possible offset, a query clip will compute $n \times h$ point-to-point distance to get one LAPD, and totally the complexity is approximately $O(nhN)$, where N is the number of clips in the training set. In this section we derive a faster differential-based algorithm, Differential Luminance Field Trajectory (DLFT). The flow chart of the proposed approach is plotted in Fig. 2.3 and we will introduce step by step in the following paragraphs.

Spatio-temporal features are extracted from the preprocessed data in the following ways: the spatial features are firstly extracted from the appearance of frames. Then a differential is applied on neighbor points of the trajectories to maintain the temporal features. In this step, the following signal is formed.

$$dx_k = \begin{cases} 0, & k\text{=1} \\ |x_k - x_{k-1}|, & \text{otherwise} \end{cases} \tag{2.13}$$

For the first frame, we skip capturing the differential information and set it to zero since there is no previous frame. From the second frame on, we record the difference between neighbor frames as a sequence signature. The information included in the signature is dominant for discrimination, but in a quite simple format. Each sequence can be represented as a group of differential signals in this way.

For a real-time application, the duration of video clips are different with each other, which means trajectories are of different length. In LAPD each possible offset is considered. In DLFT, to solve this problem and better utilize the features, a transformation is applied

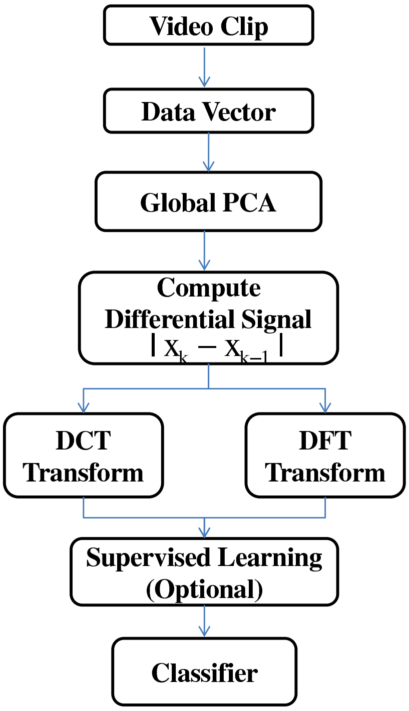


Figure 2.3. Flow chart of Differential Luminance Field Trajectory Solution

on the differential trajectory to normalize the length of the data to represent each video clip while maintaining the useful information. A 64-point DCT or DFT is performed on a sliding time window over the differential trace, as shown in Eq. 2.14.

$$\Gamma(k) \leftrightarrow \Gamma(dx_k) \tag{2.14}$$

where the $\Gamma(.)$ is the DCT or DFT transformation. Noticed that for DFT, which is not a real transform, only the magnitude of the signal is utilized as feature representation. The signals in frequency domain are extracted as features and utilized for further processing. The transformation will probably introduce one or two seconds initial delay which should be acceptable for in most real-time or on-line applications. Video sequences are equalized to same length via the transformation. After the transformation, a frequency domain subspace A_d is obtained for discrimination and then KNN classifier can be used for a final decision.

Given a query video clip q, the objective is to find a clip k in the training set which is closest to q, and label q with the label of k, i.e.,

$$q^* = \arg\min_{dx_k} ||A_d(\Gamma(dx_q) - \Gamma(dx_k))||^2 \tag{2.15}$$

where A_d is learned from the frequency domain, and dx_q and dx_k are the differential features of query and training respectively.

In this method, the computational complexity is greatly reduced: after processed by Eq. 2.13 and 2.14, a query will be compared with N clips in the training set only once to get their distance. Generally we just use 64-point transformation, so the complexity is improved to $O(N)$. The processing just keep the spatial information for the first frame and use differential from later on. Many detail in the video are lost, which may degrade the performance on recognition accuracy. There is a tradeoff between the performance an computation complexity.

2.4 DLFT and LAPD Solutions

In this section we introduce our solution of DLFT and LAPD separately. The KTH human action data set, which is tested in [44] and [72], is used as an example to illustrate the proposed method in detail.

After pre-processing by down-sample and PCA dimension reduction, some 2-D luminance field trajectory are plotted in the Fig 2.4 as example trajectories. The ones contain the "hand action", i.e., *boxing*, *handclapping* and *handwaving*, are plotted in blue, while the others containing the "body action", i.e., *jogging*, *running* and *walking* are plotted in red. From this 2-D figure we can intuitively distinguish the hand action and body action by the shape of the curves. These curves will be further demonstrated to have enough statistical information for recognition tasks.

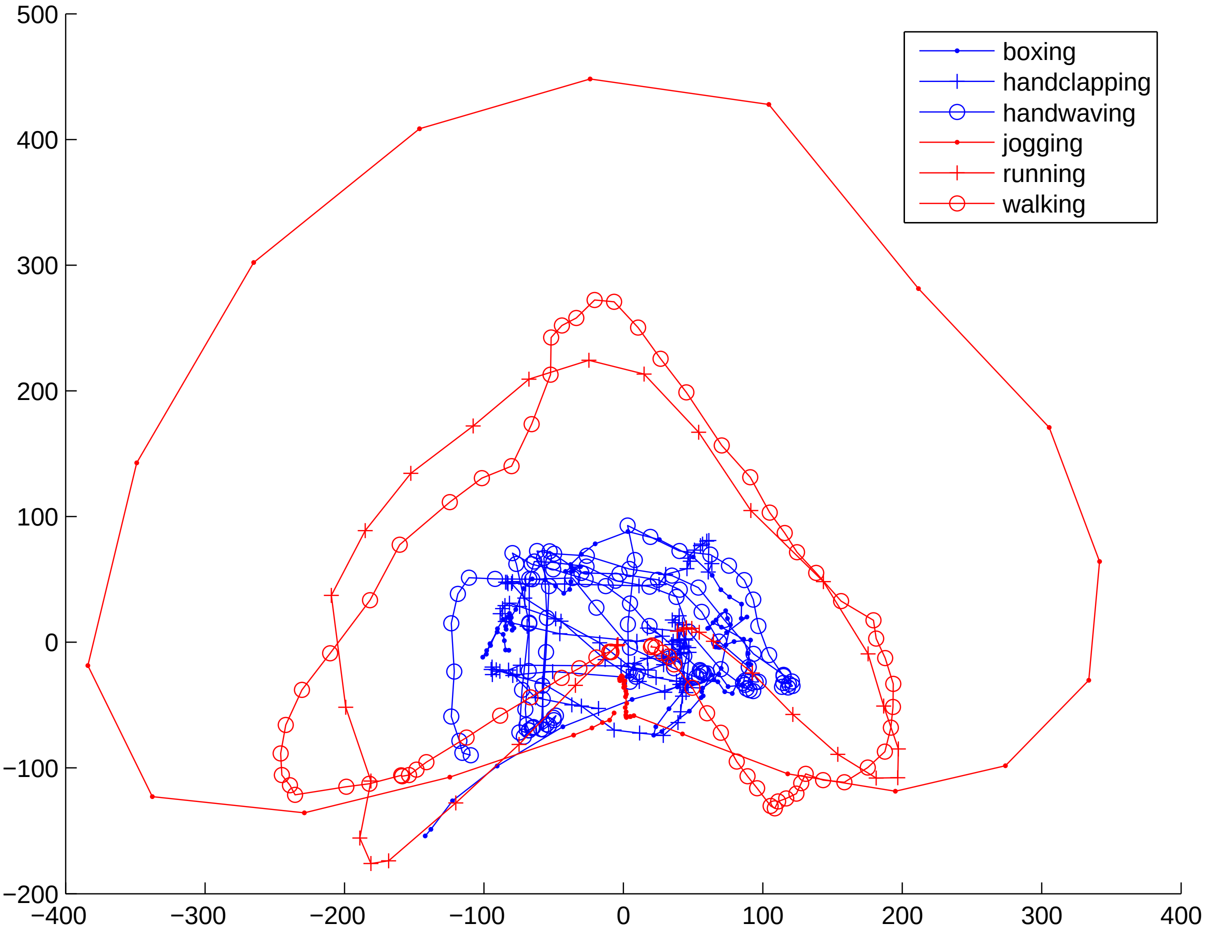


Figure 2.4. 2-D Human Action Luminance Field Trajectory Examples

2.4.1 DLFT Solutions

In DLFT solution the differential values in the trajectory are computed according to Eq. 2.13, each video clip is represented as a vector with different length. Both the spatial and temporal information is maintained in this step: the spatial information is kept in the value of signal while the temporal information are represented as the relationship between neighboring points. An example of the differential trace is plotted in Fig. 2.5. Although the shape looks different due to the different durations of sequence, it is still difficult to judge the pattern.

Then a transformation is applied to align all the clip into equal length. In this work we apply DCT and DFT separately, and the sequence after 64-point transformation is plotted in Fig. 2.6. The first several DCT coefficients are of different pattern for different kind of

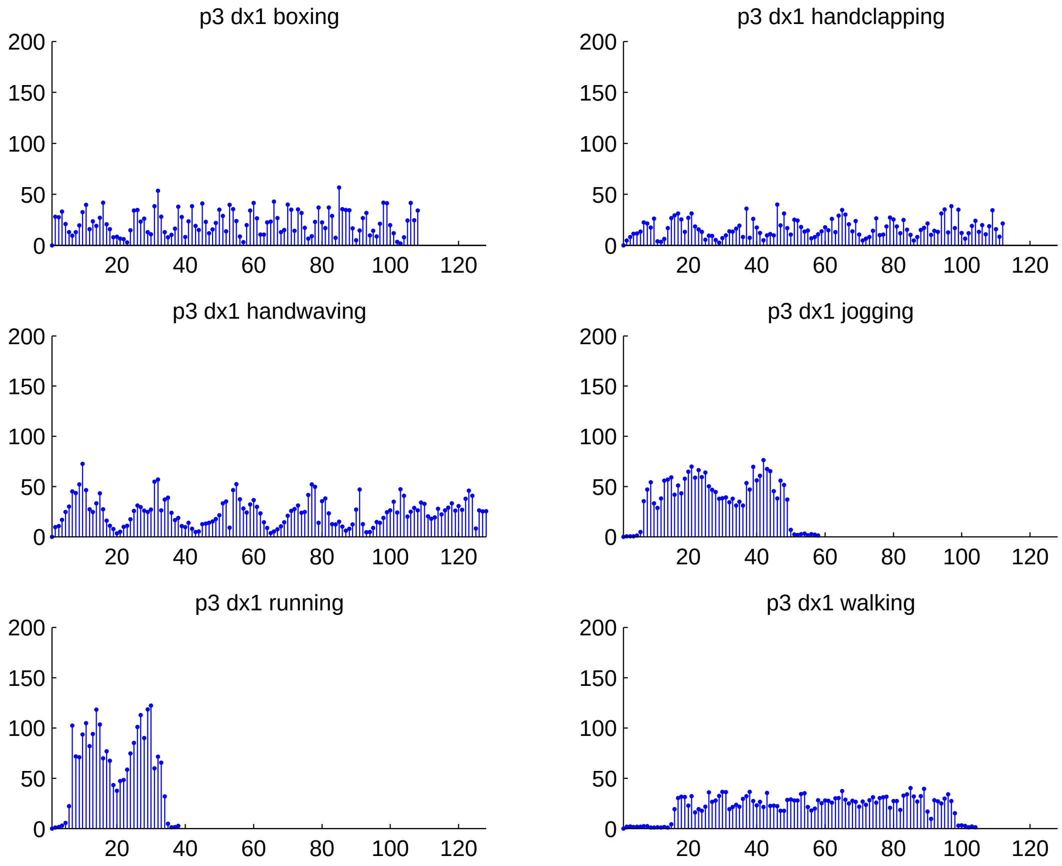


Figure 2.5. Examples for Differential Trace for Different Actions

actions, while the high-frequency coefficients are almost zero. In DFT there are also some dominant coefficients on different positions, as shown in Fig. 2.7. The certain pattern can already be understood in Fig. 2.6 and 2.7.

Standard supervised learning techniques like LDA in Fisherface [4], as well as graph embedding based Locality Preserving Projection (LPP) [31] are applied to these feature vectors in the DCT or DFT domain.

The classification is based on a simple K Nearest Neighbor (KNN) classifier. The training clips projected in LPP or LDA subspace are modeled as a 6-class Gaussian mixture, and the classification is done by assigning maximum likelihood action labels. Experimental results and discussions are presented in Section 2.5.

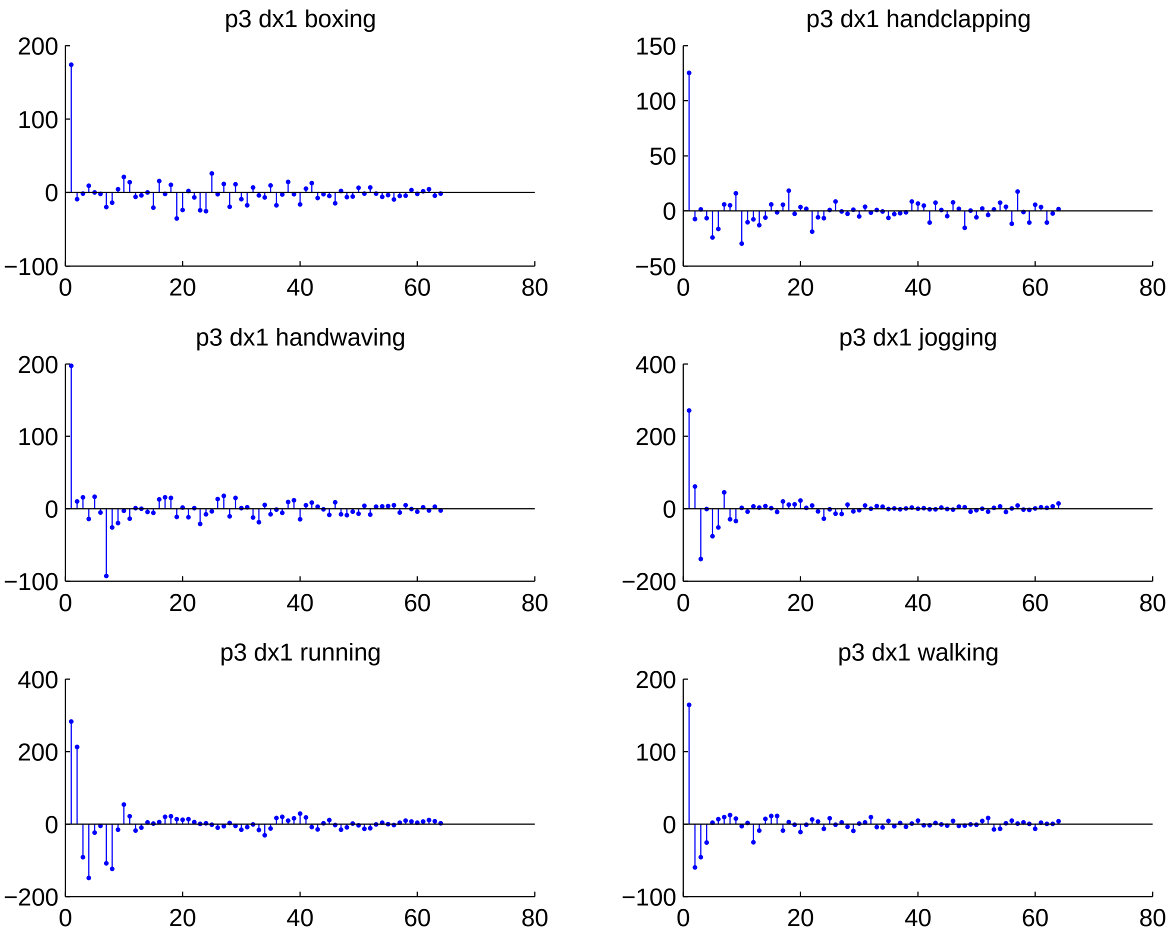


Figure 2.6. Transformed value of differential traces by DCT

2.4.2 LAPD Solutions

In the LAPD processing we focused on the computation of inter-trajectory distance. The trajectories are the same as plotted in Fig. 2.4.

In this solution, given a query clip, the distance between the query and each training trajectory is pairwise computed by Eq. 2.13, and a distance array is generated for each query. The distance from the query to each action category is computed and a histogram is shown in Fig. 2.8. The decision is made by selecting the smallest distance, which is quite obvious in the figure.

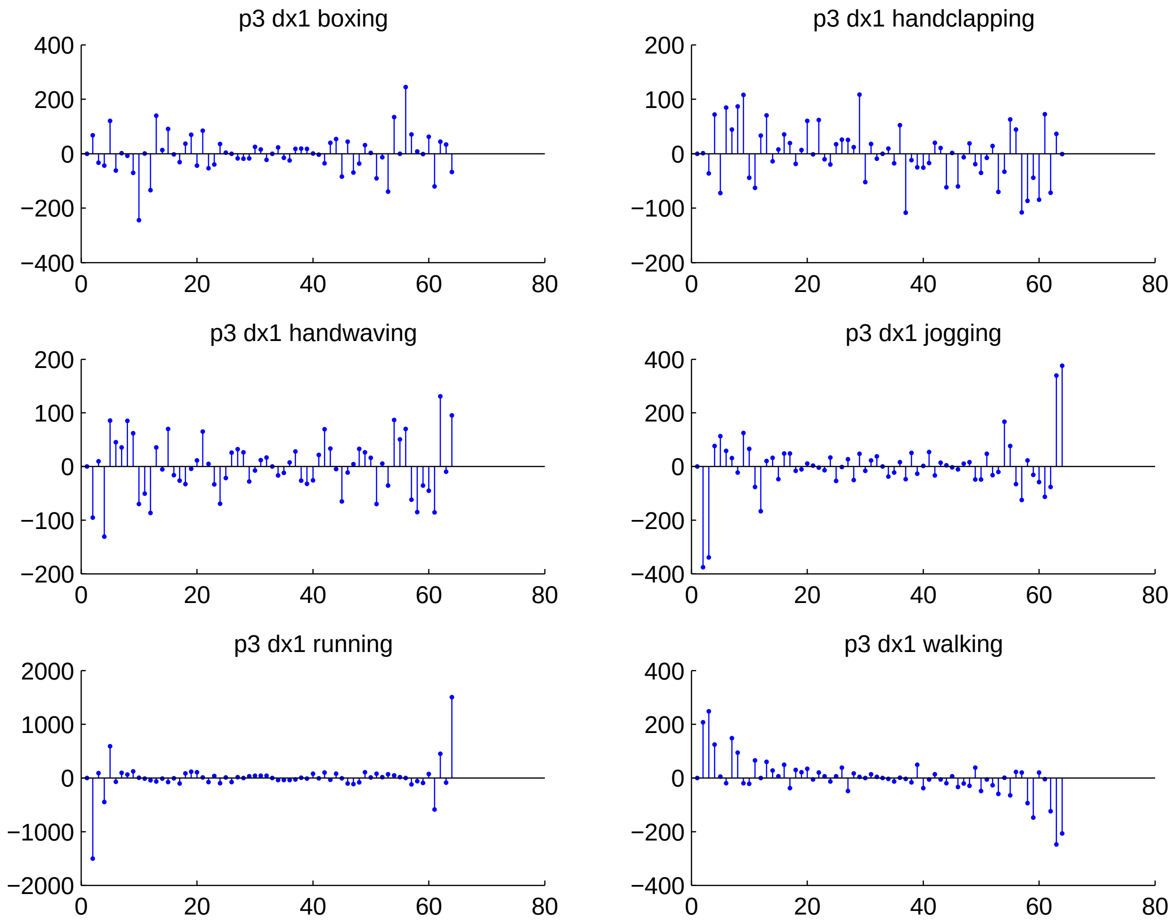


Figure 2.7. Transformed value of differential traces by DFT.

2.5 Experiments and Performance Evaluation

We have tested our methods on three different datasets, the *KTH human action dataset*, the *Cambridge Hand Gesture dataset* and the *Youtube action dataset* for human action recognition and hand gesture recognition separately. These datasets cover variations in appearance, illumination, background and spatio-temporal cues. Besides proposed approaches, we also tested the Dynamic Time Warping method on the dataset above, and compare with some results obtained by other approaches in the recent literature.

For all the datasets, our implementation is based on the leave-one-actor-out setting, where the classifier is trained using all video sequences except those corresponding to the actor in the test video. This processing is repeated many times until each video has been

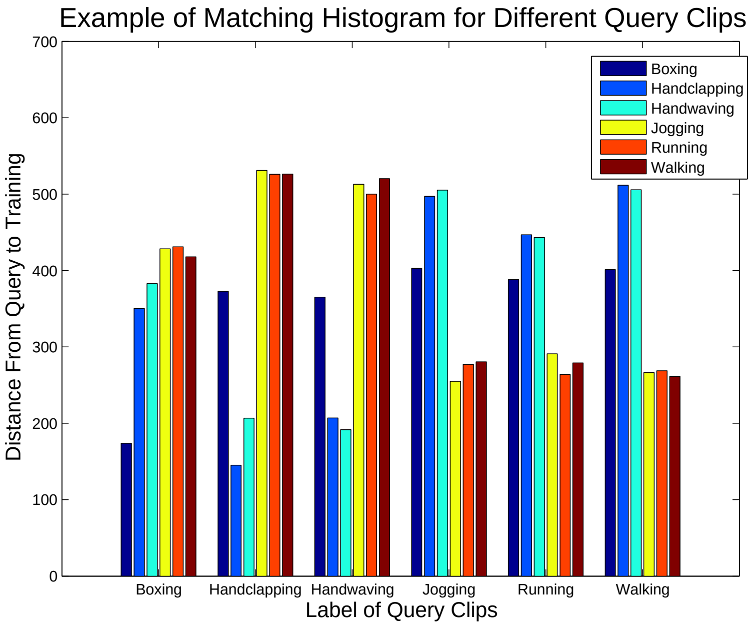


Figure 2.8. Examples for Differential Trace for Different Actions

treated as the test video.

2.5.1 Recognition Accuracy

2.5.1.1 KTH Human Action Dataset

To test the developed algorithm, we use the human activity data set from [72], which contains 6 human actions, *'boxing'*, *'handclapping'*, *'handwaving'*, *'jogging'*, *'running'*, and *'walking'*. Actions are performed by a total of 25 subjects in 4 different settings:

S1: outdoors;

S2: outdoors, with camera zooming;

S3: outdoors, with different clothes on;

S4: indoor.

For each setting, each action has 4 video clips, with each segment's start and end frame number listed as a ground truth file. Each setting will have $4 \times 25 \times 6 = 600$ actions, and the data set comprises of a total of $2,391$ clips, with a small number of entries missing.

The video clips are of 160×120 pixel resolution, and in processing stage, we down convert the sequence into 20×15 icon image sequences for trajectory modeling. Some examples from the action clips in [72] are plotted in Fig. 2.9. From left to right, the top row actions are, walking, jogging, and running, and the bottom row actions are, boxing, hand waving and hand clapping.

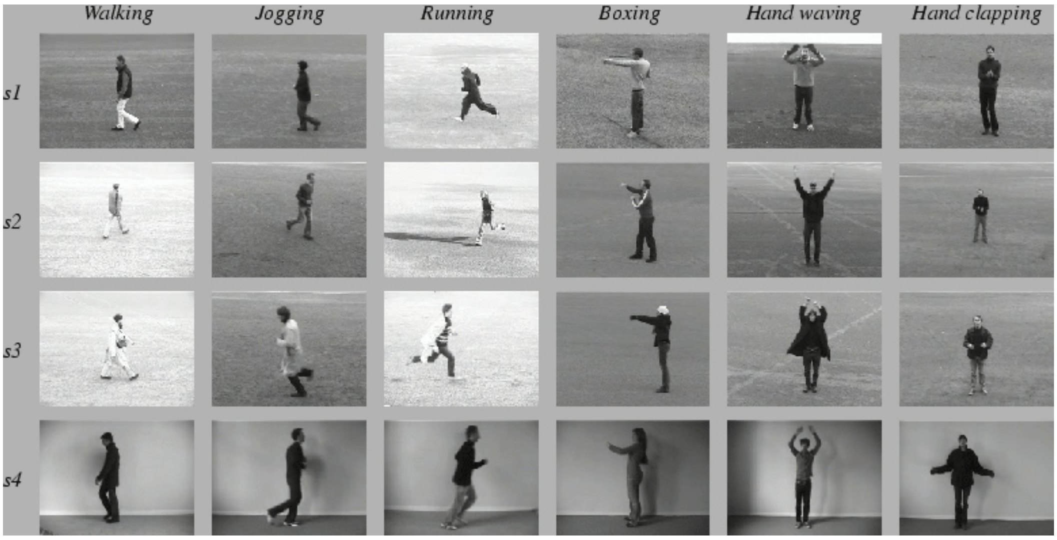
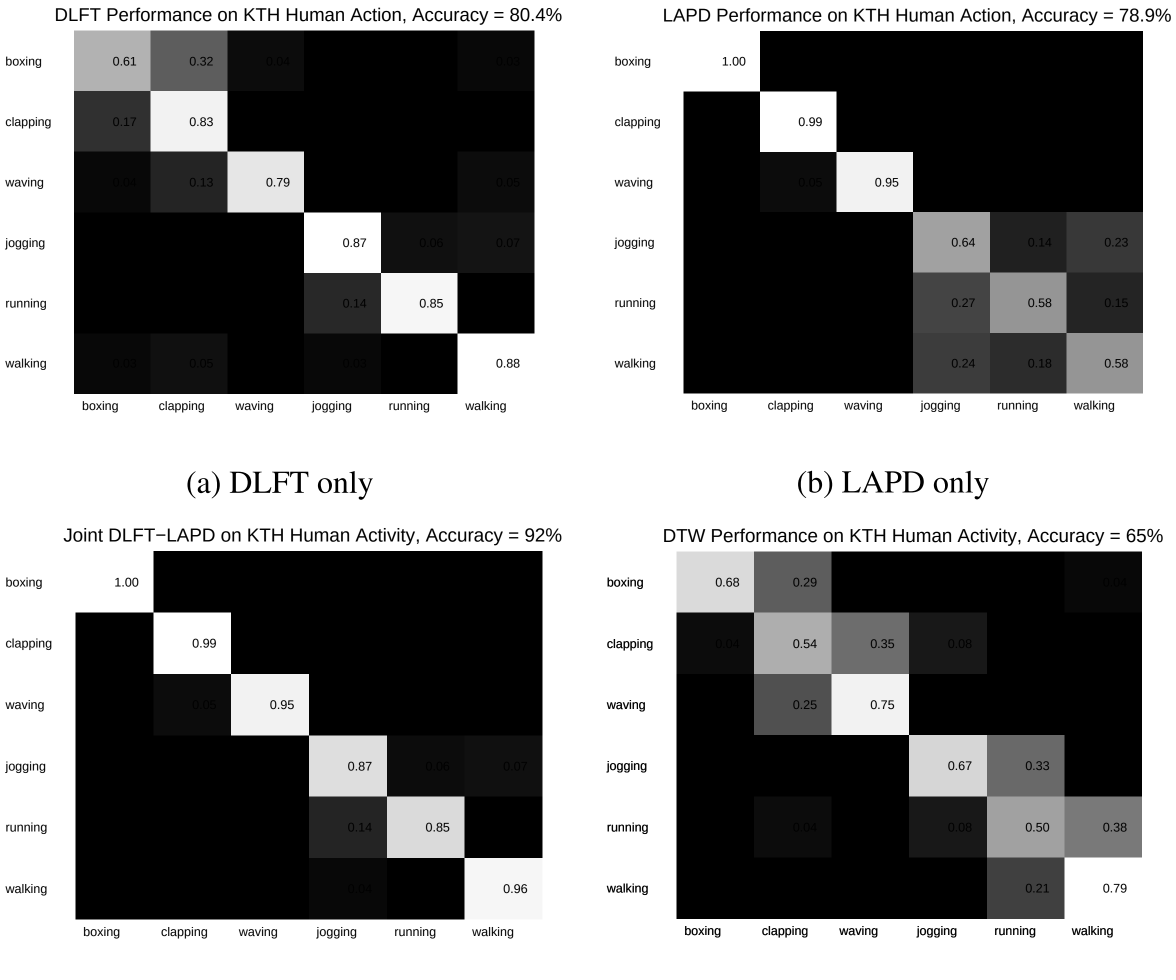


Figure 2.9. Sample frames in KTH human action dataset [44]

In the pre-processing stage, we down-sampled the original video frame to 20×15 icons and applied a global PCA to reduce the number of dimensionality to 64. Thus the video sequences are represented by trajectories in $\Re^{64}$ and the differential trajectories are computed. To deal with the different durations of video clips and eliminate the high-frequency coefficients, a 64-point DFT is taken. Then LDA supervised learning is applied on the differential signals and KNN classifier is used to assign a label for the test video sequence. The confusion map is shown in Fig. 2.10(a), and the overall recognition accuracy is 80.3%

For the LAPD approach, the 64-dimensional feature is used again after pre-processing for dimensionality reduction. Supervised learning is later applied to discriminate the differ-



(a) DLFT only

(b) LAPD only

(c) Joint DLFT and LAPD

(d) DTW

Figure 2.10. Performance on KTH human action recognition(a) by DLFT only, (b) by LAPD only, (c) Joint DLFT and LAPD, (d) DTW.

ent patterns and to further reduce the number of dimension for easier LAPD computation. The result is shown in Fig. 2.10(b), with an overall accuracy of 78.9%.

From the result of the proposed two approaches, it is observed that DLFT has better performance on behaviors containing more temporal information, such as *running* and *jogging*, while LAPD has advantages in appearance-based behaviors like *handwaving* and *handclapping*. Intuitively the performance can be further improved by combining the strong points of both methods. During the test it is found that differential trajectory of the first 3 behaviors, i.e., *boxing*, *handwaving* and *clapping* is proved to be less sparse than the other 3 actions, i.e., *jogging*, *running* and *walking*. A combination scenario is therefore implemented based on this observation.

The combination is processed with the following steps. In the feature extraction step, both the luminance trajectories and their differential values are stored. Then the strength of the temporal information is evaluated by measuring the sparseness of the differential signals. A threshold is set to discriminate whether the temporal features are stronger than the spatial ones or not. When there is stronger temporal information, it means there is large motion in the video sequence and DLFT is applied to classify the query sequence into the last 3 actions, i.e., *jogging*, *running* and *walking*. Otherwise, LAPD is applied to utilize the appearance features for discrimination among *boxing*, *handwaving* and *clapping*. The result of the combination of LAPD and DLFT is shown in Fig. 2.10(c) with a final accuracy of 92%.

For better comparison, the DTW algorithm is also implemented in our experiment. With the same pre-processing, the DTW is applied directly on the trajectory in $\Re^d$. An optimal alignment can be found between two trajectories and the point-to-point Euclidean distance is computed as a measurement of similarity. The label decision is done with a KNN classifier and the result is shown as in Fig. 2.10(d). The overall accuracy is 65%.

2.5.1.2 Cambridge Hand Gesture Dataset

To demonstrate the robustness and versatility of the algorithm, we also tested another dataset, the *Cambridge Hand Gesture Dataset*, which is composed of 900 image sequences with 9 different hand gesture classes [42]. These classes are defined by 3 primitive hand shape: *Flat (F)*, *Spread (S)* and *V-shape (V)*, and 3 primitive motion directions: *Leftward (L)*, *Rightward (R)* and *Contract (C)*. There are totally 9 gesture classes by combining the two factors above. Each class contains 100 image sequences, with 5 different illumination cases. An example is shown in Fig. 2.11.

Each sequence in the dataset has a different number of images, varying from 37 to 119, and the total number of image is $63,188$. The original image is a 320×240 color image, and we firstly reduce the number of data by converting it to gray image and then down-sample it to a 32×24 icon. We then process these icons with the DLFT and LAPD

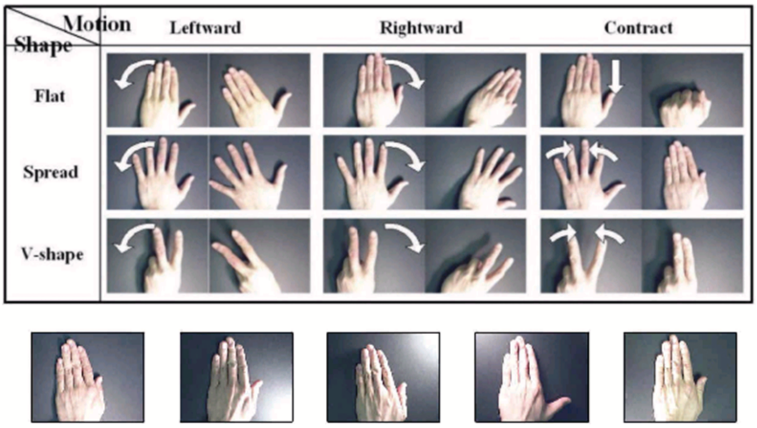


Figure 2.11. Sample frames in Cambridge hand gesture dataset, (top) 9 gesture classes by 3 motion directions and 3 hand shapes; (bottom) 5 different backgrounds with different illuminations [42]

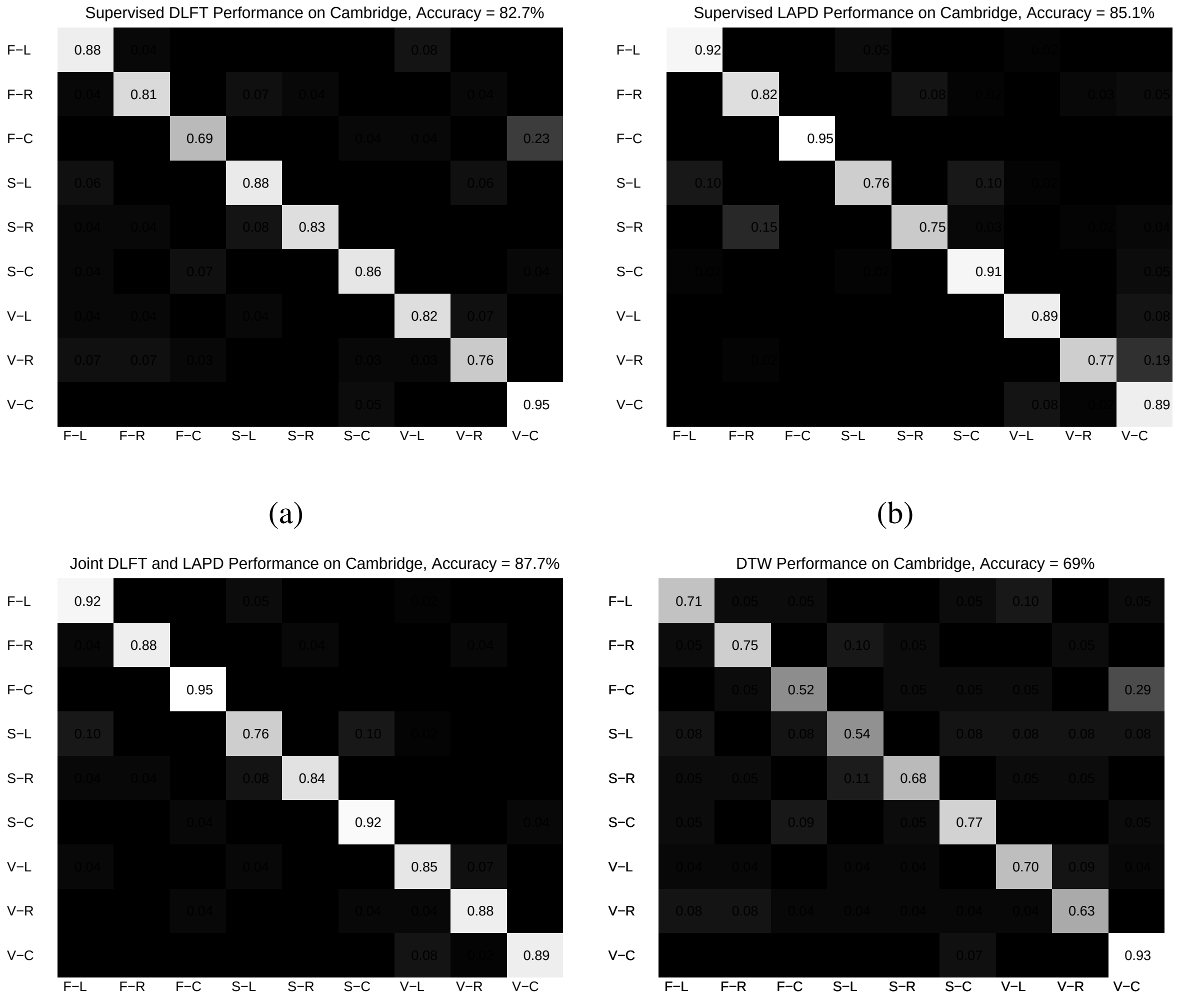
approach, respectively.

The hand is composed by many connected parts and the motion is highly articulated. Without prior information, it is difficult to guess what kind of appearance is contained in the image even if it is known to be a hand. To better discriminate the appearance, we keep 32×24 pixels as icons in the preprocessing. For a global dimensionality reduction, a PCA is applied on $\Re^{32\times24}$ and 64-dimentional trajectories are treated as representations for the image sequences for further processing.

With the DLFT approach, a 64 point DFT is applied directly on the trajectory to solve the duration problem. The simulation result is shown in Fig. 2.12(a), the overall recognition accuracy is 82.7%.

In the LAPD approach, the distance is computed and the label of the query clip can be decided by a KNN classifier. The recognition accuracy is 85.1%. Figure 2.12(b) shows the confusion map for gesture recognition. For the individual set testing, the comparison results are listed in Table 2.1.The numerical result is comparative to the one in [42] and better than some other results reported in the literature.

In order to better utilize the correlation between spatial and temporal features in the



(a)

(b)

(c)

(d)

Figure 2.12. Performance on cambridge hand gesture recognition(a) by DLFT, (b) by LAPD, (c) by Joint DLFT+LAPD, (d)by DTW.

video sequence, a joint of DLFT and LAPD is applied for hand gesture recognition. The result is shown in Fig. 2.12(c), with an accuracy of 87.7%.

Dynamic Time Warping has also been tested on the hand gesture dataset and the result is shown in Fig. 2.12(d). The accuracy is not as high as with LAPD, since even in the same hand motion pattern, the magnitude of motion may be different with different people.

The object of appearance information in the hand gesture data set is mainly the hand, and the change in the background is a factor to test the robustness of either proposed algorithm. From our KNN result, we found out that in the correct cases, the nearest trajectory in

Table 2.1. Hand Gesture Recognition Accuracy Comparison(%).

Method	Set1	Set2	Set3	Set4	Set5	Average
DLFT(Unsupervised)	76	75	68	74	72	73.0
LAPD(Unsupervised)	83	81	77	80	79	80.0
DLFT(Supervised)	84	81	86	84	78	82.7
LAPD(Supervised)	91	85	78	84	87	85.1
TCCA [42]	81	81	78	86	–	81.5
Nieble [58]	70	57	68	71	–	66

the training set is always in the same class and same background as the query clip. The effect of the changing of the illumination of the background will not influence the recognition performance.

As shown in both [42] and our LAPD solution, there are several confusions between the class *spread* and *flat*, for either left or right direction. These are mainly due to very little difference in appearance, and the details are lost when sampling the original frames down to a small icon. The DLFT approach, however, is not suffering from this problem, because the differential operation cares more about the temporal properties, compared with the spatial one. Therfore, a joint of the DLFT and LAPD can take advantage from both spatial domain and temporal domain, and the recognition accuracy is better than the separate method.

2.5.1.3 Youtube Action Dataset

A more challenge dataset we used to test our proposed solution is the *Youtube human action dataset* [50]. 11 different actions, *Basketball shooting*, *Cycling*, *Golf Swing*, *Diving*, *Horse Riding*, *Soccer Jungling*, *Swinging*, *Tennis Swinging*, *Trampoline Jumping*, *Volleyball Spiking* and *Walking with a Dog*, are included in the dataset. Each action is repeated by 25 different persons for several times and totally $1,577$ video clips are involved. Example frames are shown in Fig. 2.13.

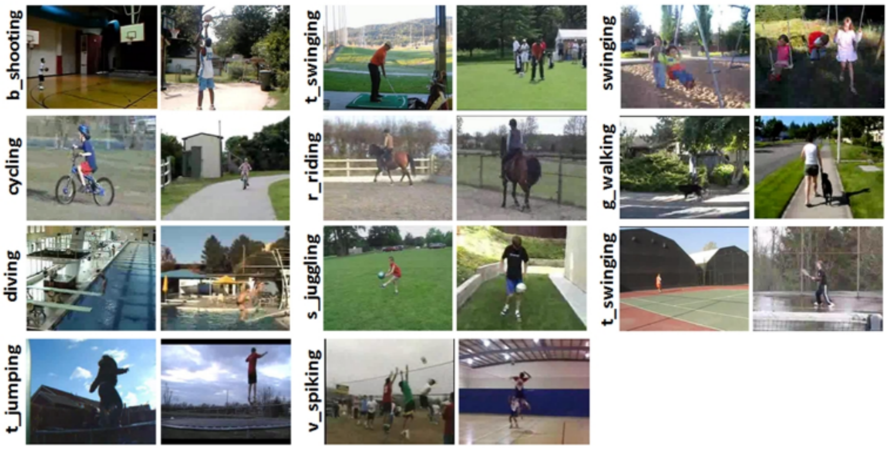
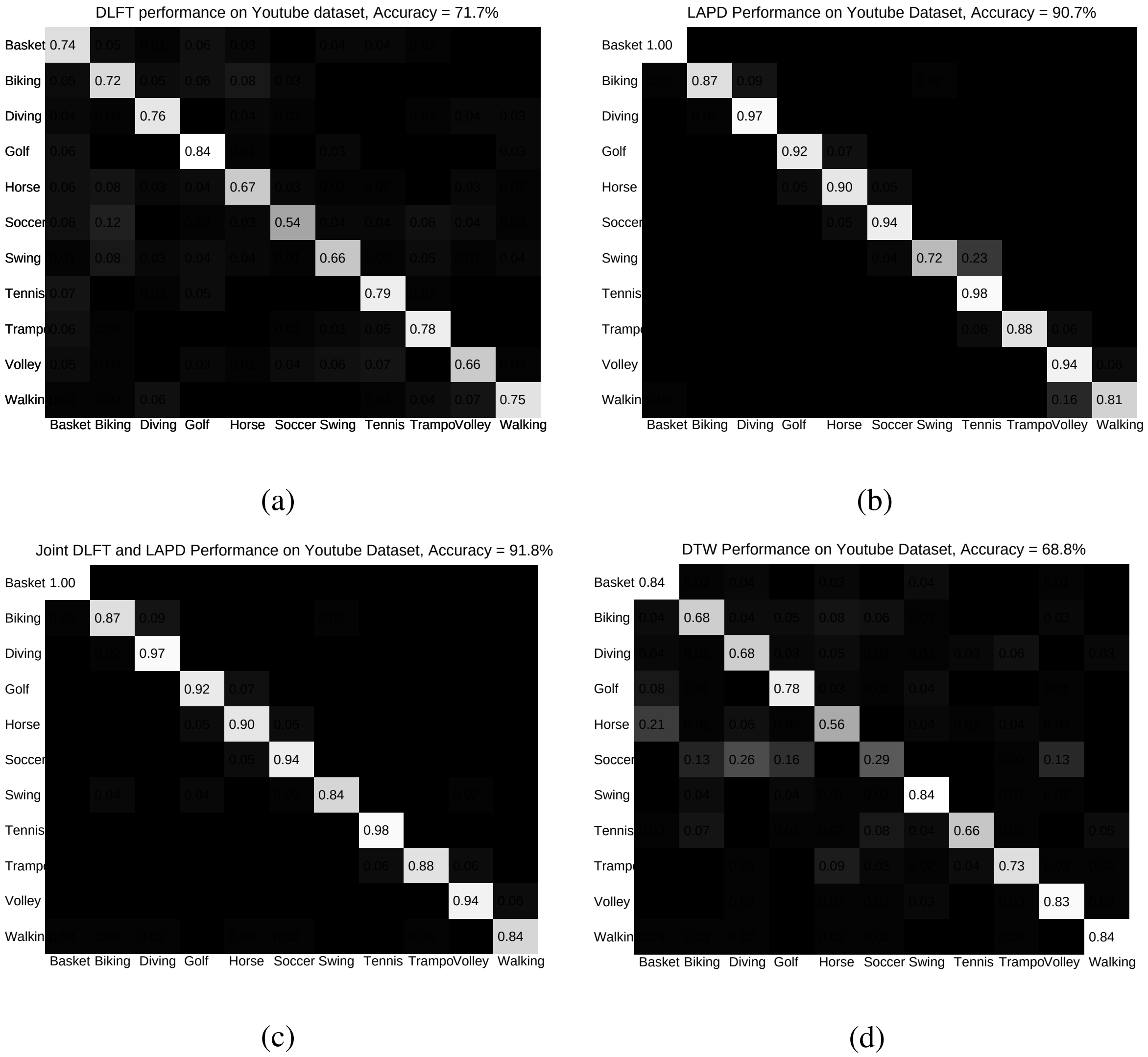


Figure 2.13. Sample frames in Youtube dataset, [50]

This is a challenge human action dataset because of large variations in camera motion, object appearance and pose, object scaling, viewpoint and very complicated background. Therefore, in our DLFT approach, we applied a 64-point DCT instead of the DFT in the proposed algorithm. The result of the DLFT method on the Youtube database is shown in Fig. 2.14(a). The recognition accuracy is 71.7% for all eleven actions.

The first 64-dimensional feature was kept in the LAPD and DTW experiment. A global subspace is learned with the training set. The results are shown in Fig. 2.14(b) and Fig. 2.14(d) respectively. LAPD results in an accuracy of 91% and DTW in about 65%. A combination of LAPD and DLFT is also applied on this set, with the result plotted in Fig. 2.14(c).

Since there are very complicated background variation in the video sequence, the temporal features does not contribute as much as spatial one. The result from our proposed DLFT method is comparable with the ones in the literature such as [50] and [51], but not as good as LAPD.


DLFT performance on Youtube dataset, Accuracy = 71.7%

LAPD Performance on Youtube Dataset, Accuracy = 90.7%

Joint DLFT and LAPD Performance on Youtube Dataset, Accuracy = 91.8%

DTW Performance on Youtube Dataset, Accuracy = 68.8%

(a)

(b)

(c)

(d)

Figure 2.14. Simulation result for Youtube Action dataset: (a) DLFT approach (b) LAPD approach (c) Joint DLFT-LAPD approach (d) DTW algorithm.

2.5.2 Parameter Selection in Experiments

In the experiment, there are some degree of freedom to choose parameter settings. Some of the settings will affect the performance of proposed method. In this section, we present our selection of parameter and give corresponding analysis on the reason to select.

2.5.2.1 Resolution and Dimensionality

As presented in Fig. 2.1, the preprocessing of the video sequence is composed by two sequential steps, down sample and PCA. In this section we will analyze the related parameter selection in these two steps.

The first step is down sample. Given the original video frame with resolution $W \times H$, the preprocessing down sample it to a smaller icon $w \times h$. In our experiment, we find that the resolution of icon will not greatly affect the recognition accuracy, but larger icon will take longer time for processing. Therefore, according to the size of each dataset, we choose the minimal icon size that can maintain the statistical information. In the KTH human action dataset, we are using 20×15 icon. In the other two dataset, Cambridge hand gesture dataset and Youtube dataset, we are using icons with resolution 32×24 due to more detail required.

The second step in preprocessing is PCA for dimensionality reduction. This operation is a tradeoff between the speed and accuracy. More dimension will preserve more useful information for discrimination, but slow down the recognition processing as well. The performance will also be drastically degraded if there is no sufficient features. Therefore a proper number of dimension is very important for recognition performance. In our experiment we are using the following approach, during PCA we have the eigenvalue for each dimension in subspace. The importance of dimension is reflected by the corresponding eigenvalue. Eigenvectors with larger eigenvalue will have more contribution in discrimination. So we sort the eigenvalues in descending order and preserve the first k dimension.

The first k eigenvector contains the most important statistical information in recognition, and in Fig. 2.15 we plot how many percentage of information is preserved of KTH action dataset as an example. The percentage is defined as follow,

$$\rho(k) = \frac{\sum\limits_{n=1}^{k} \lambda_n}{\sum\limits_{n=1}^{N} \lambda_n} \tag{2.16}$$

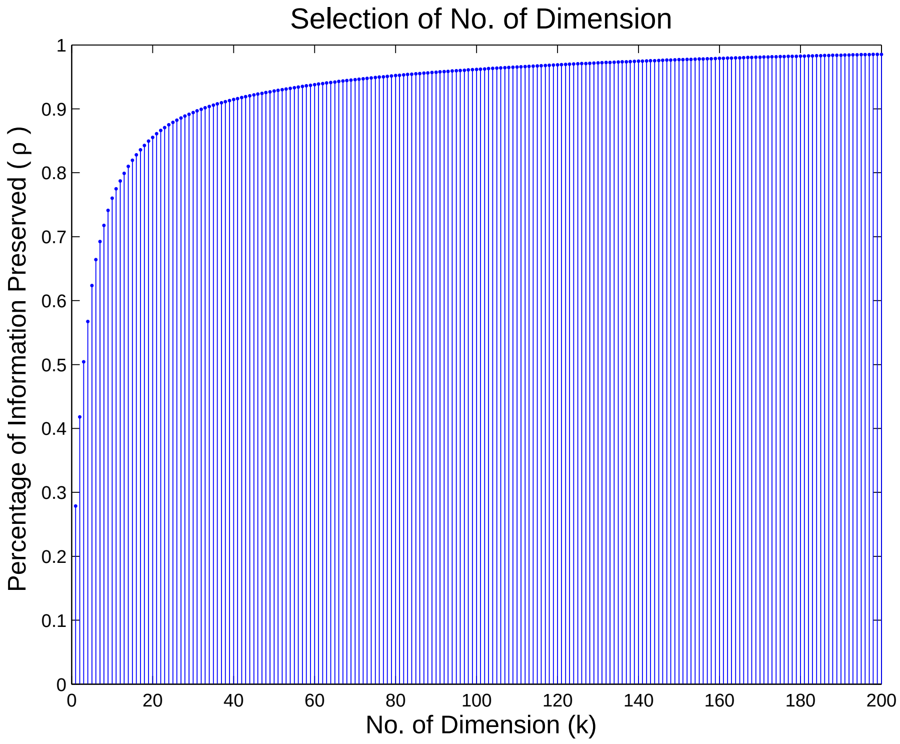


Figure 2.15. Parameter selection in PCA for dimensionality reduction

Table 2.2. Selection of number of dimensionality.

Number of Dim.	Percentage (%)
16	83.1
32	89.9
64	95.1
128	97.0
256	99.4

where λ_n is the nth eigenvalue in PCA computation, and totally there are N eigenvalues. We summarize some important value in Fig. 2.15 and show it in Table 2.2,

From table we can see that more than 95% of information is preserved in the PCA

Table 2.3. Parameter selection in KNN classifier.

Value of k.	Recognition Accuracy (%)
1	73.3
3	75.0
5	78.9
7	77.8
9	75.0

when 64-D features are extracted, which is already enough to achieve a satisfactory recognition accuracy. Therefore in our experiment, we finally reduce the number of dimensionality to 64. Similarly, in the Cambridge hand gesture dataset and Youtube dataset, we adopt the same setting for dimensionality reduction.

2.5.2.2 Classifier Parameters

As presented in previous sections, we are using KNN classifiers to decide the query label in recognition in our proposed LAPD approach. In the classifier, k nearest neighbors are used as reference, so the number of k is a parameter which can be freely set in the experiment. Too small k value will cause unreliable recognition result, while too large k will result in unrelated samples in the nearest neighbor set. Therefore, a proper k value is necessary to achieve high recognition accuracy.

Since the query is classified by a majority vote of its neighbors, the number of k is better to be odd. In 2.3, we list the possible values of k and their corresponding recognition accuracy on KTH human action dataset.

From the table we can see that the performance is slightly different. The performance is best when the value of k is set to 5. Therefore we select $k = 5$ as the parameter for KNN classifier. The result is similar in other dataset, so to be fair, in our experiment we are all

Table 2.4. Parameter selection in DLFT Approach.

Trans Length	DCT Accuracy (%)	DFT Accuracy(%)
16	71.3	72.3
32	73.0	73.3
64	75.0	76.5
128	72.5	74.0

using $k = 5$ for the classifier setting, in all the dataset.

2.5.2.3 Parameters in DLFT

Compared with LAPD approach, DLFT has more flexibility. The flow chart of processing is shown in Fig. 2.3. As presented in previous sections, we are using 64-Dimensional feature after global PCA, and finally the KNN classifier is applied with parameter k set to 5. Besides, we need to process the differential signal with a transform, either DFT or DCT, with different possible choice of parameters.

The selection of number of points for transform is a balance between the computation and accuracy. Smaller points of transform will result in a faster speed, but a lower accuracy. On the other hand, when the number of points are very large, it will go beyond the length of differential trajectory, which may introduce noises in transformation.

We are using KTH dataset to show the progress of parameter selection. Recognition accuracy is listed in Table 2.4.

From the table we can see the 64 point transformation gives best recognition accuracy. So in our experiment we uniformly set the transformation length to 64. For different transformations, DFT provide better recognition performance than DCT, on various transformation length, for about 1 to 2 percent.

2.5.3 Result Analysis and Discussion

The proposed DLFT and LAPD approaches are mainly aiming at keeping both spatial and temporal features, which are crucial for action recognition in video sequences. In the KTH dataset, the 6 classes of human action have different spatio-temporal content: the three hand actions have more spatial information than the temporal one, while the body action mainly consists of temporal features.

For LAPD, the dominant feature for recognition is the distribution of frames in the transformation domain. The aligned method is designed to address the difference in video durations. As is proved in the experiments the performance is quite accurate, even with a small number of training samples.

The DLFT is mainly focused on the difference between neighboring frames by utilizing the differential value. The statistical feature, which is composed for classification, is the variation distribution of differences in the video sequence. This feature does not consume many resources due to the limited number of video frames, and can be used for on-line applications. It is also theoretically and practically robust to the change of background or occlusions, because after taking differences the unstable factors above are all removed.

Both methods above still have a common advantage: they are both content independent and only very simple pre-processing is needed. They can be applied to both human action recognition and hand gesture recognition. Theoretically the algorithm itself, is not dependent on the video content, and for every possible video sequence, the method can offer good recognition accuracy as long as the classes are defined clearly. Generally, LAPD paid more effort on appearance modeling, with higher accuracy, while DLFT focused mainly on temporal modeling, with a faster response.

2.6 Summary

In this chapter, we proposed two approaches for video pattern recognition without object level learning. The DLFT solution is computationally efficient. The major time cost is brought by the using of the DCT or DFT. Generally, such one or two seconds can be acceptable for most real-time applications, especially large scale scenarios.

Another solution we proposed is to utilize aligned projection distance approach. We still use the trajectory to represent the video clip and calculate the distance between every two clips to define how similar they are to each other. The results show that the recognition accuracy is comparable or better than other techniques in the literature. Regarding complexity, the on-line recognition for the query is very fast and is also suitable for most applications when the training set is fixed. With increasing size of training dataset, the timing complexity is growing linearly, which is acceptable for most applications.

CHAPTER 3

LOCALIZED SPATIO-TEMPORAL INDEXING WITH DYNAMIC WARPING FOR HUMAN ACTION RECOGNITION

3.1 Overview

In the previous chapter we proposed a global subspace learning approach, luminance aligned projection distance, to solve the human action recognition problem. The video sequences are represented as trajectories in the high-dimensional space, spatial and temporal statistic information are utilized to find the similarity between such trajectories. However, details of video clips are skipped in the dimension reduction process and did not contribute to the recognition task. In other words, the local information is not emphasized in the learning process, the appearance modeling is not strong enough since the discrimination is mainly achieved by global features.

In this chapter we focus on improving the recognition accuracy by utilizing the local features in the video clip. We develop a localized appearance modeling and indexing scheme and apply it for human action and gesture recognition. The proposed approach allows querying video action clips to identify a localized appearance neighborhood, which is more accurate compared with the global solution. These local appearance metrics are then used for computing the aligned projection along the temporal trajectory. In this way, the similarity between trajectories will be refined to a higher accuracy, and the distance metric is becoming more reliable.

The local information in the video sequences can be described in various kinds of representations. In [31] Locality Preserving Projections (LPP) is proposed to describe the spatial and temporal neighbor relationship by constructing an affinity matrix. In [101] splined interpolation is applied to simplify the computation, and graph embedding is utilized to describe the relationship between neighbors. In [42], Tensor-based Canonical Correlation Analysis (TCCA) is proposed, local information is represented by correlation in this approach. Local appearance information is considered jointly with temporal order, to get a subspace for dimension reduction.

In this chapter we will use the same video representation processing steps as in the previous chapter. We describe the local information by utilizing statistical information to split the appearance space into many sub-blocks. Every sub-block in the approach can be treated as a local subset of the whole appearance space. In the learning stage, a local subspace will be learned for each subset. The similarity between trajectories will be evaluated by computing local aligned projection distance, instead of global one.

Flexibility of the solution is also improved in this chapter. In the previous LAPD approach we are doing one-by-one matching along the time direction, i.e., if kth frame in one sequence is matched with lth frame in another sequence during the alignment processing, the $k+1$th frame will be certainly matched with $l+1$th without any flexibility. However, this may not always be true in practise. It is possible that people are doing a same action with different pace and in different session.

To address this problem, the idea of Dynamic Time Warping (DTW) [38] is introduced in our approach. For inter-trajectory point-to-point matching, we allow a few point flexibilities for every single frame. The affect brought by temporal information is enhanced in this processing and matching will become more reliable in the alignment computation.

The rest of this chapter is organized as follows. We will briefly introduce the problem into a local version and the procedure for appearance space splitting and indexing in section 3.2. In section 3.3 we will present the spatio-temporal modeling together with the query driven alignment matching solution. In section 3.4, warping-based method is proposed to better solve the problem. The experiment result will be presented in section 3.5 and finally a summary session will be in section 3.6.

3.2 Local Appearance Modeling

In this section we are dealing with the modeling problem for local appearance space. In our previous global subspace learning, the detail which exists in the local appearance space is neglected in proposed approaches. In order to utilize such local information, we differentiate different area in the appearance space with different subspace parameters.

3.2.1 Preprocessing

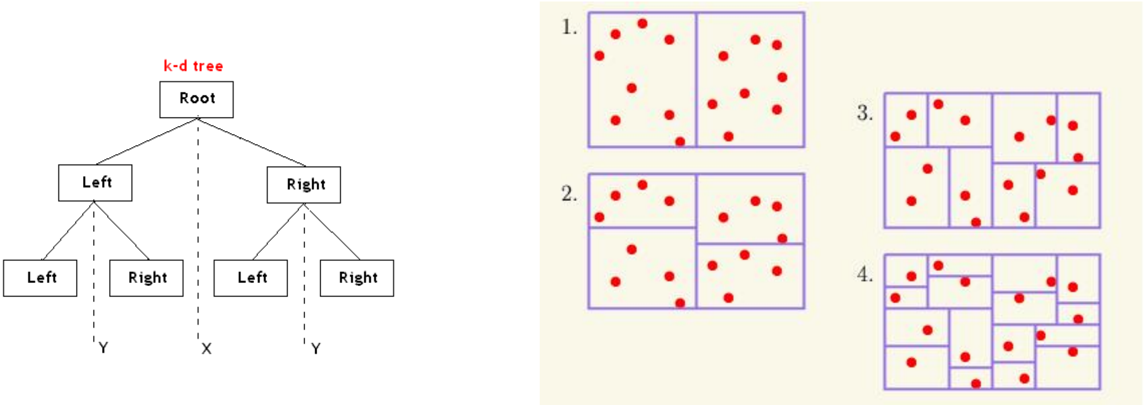
In the preprocessing stage we convert a video sequence into a trajectory in a high dimensional space. The preprocessing in this chapter follows the one in the Chapter 3, and we briefly repeat the technique as follow.

The preprocessing starts from converting each video frame into a vector. Originally each frame is represented as a RGB metric with an M by N by 3, where M by N is the original resolution. Since the data is too large to efficient processing, the frame is firstly converted to gray and then a downsample is applied to reduce the original image down to a small icon, with size m by n. Then the small icon is resized as a vector as 1 by mn, therefore a video sequence with k frames can be represented as a matrix k by mn.

A global PCA is applied to further reduce the number of dimension in the video processing. The number of dimension per frame will be reduced down to d after the computation. Up to now, one single video sequence with k frames is represented as a k by d matrix, and therefore a dataset composed with N video frames will become N vectors in the d dimensional space.

3.2.2 Spatio-Temporal Appearance Indexing

As the training data set grows in size, it is a great challenge to model this local appearance information. In this work, we propose a solution to handle the problem by partitioning the appearance space with an indexing scheme in order to collect the similarity within a single appearance block. The partitioning is conducted level by level into a K-D tree structure, as shown in Fig. 3.1. In each partitioning the number of points in one branch is equal to that in the other, i.e., half points are clustered into one branch and the others into the second branch during every partitioning.



(a) K-d tree structure

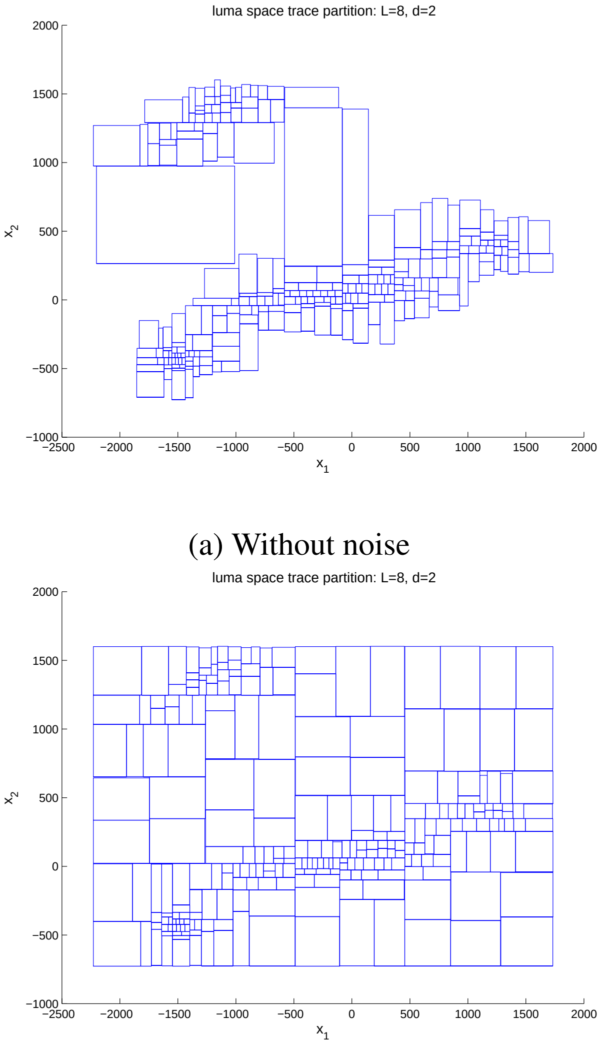
(b) Theoretical Partitioning

Figure 3.1. K-D tree structure and space partitioning [22]

At each splitting, the clustering starts by computing the variation in each dimension. In order to enhance the effect of splitting, we focus on dividing the dimension with maximum discrimination every time. In each round every leaf node in the last level of the tree will split into two new leaf nodes, and thus the level of the tree is increased by one. This so-called maximum-variation splitting offers an optimal partitioning of the frame field, and the appearance information can be effectively utilized for discrimination. The tree will finally become an L-level binary tree by iteratively splitting for times in $2^L - 1$ rounds.

In this method, the boundary of the partitioning is defined by the minimal and maximal bounding box (MBB). All boundaries are straight lines instead of curves with various shapes. Given a set of data $\{x_1, x_2, ..., x_n\} \in R^d$, the minimal and maximal value for each dimension j of the set can be obtained by finding:

$$\begin{aligned} lb_j &= \min(x_1^j, x_2^j, ..., x_n^j); \\ ub_j &= \max(x_1^j, x_2^j, ..., x_n^j); \end{aligned} \tag{3.1}$$



(a) Without noise

(b) With white noise

Figure 3.2. 2-D Appearance space partitioning example with L=8 (a) Original frames; (b) With additive white noise

where lb_j and ub_j are the lower bound and upper bound of the j^{th} dimension in R^d and x_k^j denotes the j^{th} dimension of the data x_k.

As an example, when $d = 2$, the MBB will be a rectangule. Figure 3.2(a) gives an example of the binary tree with $L = 8$ in the 2-D case. However, the appearance space is sparsely populated due to the limited training samples in the dataset. Several leaf nodes are occupying a very large volume and make the indexing less efficient. This has implications in the query-driven local appearance subspace training.

To address this problem, we introduce random noise points in the space. The upper and lower bounds of the overall data are calculated and then random noise points with uniform distribution are added. It guarantees that each partition is not very large. The noise points are only used for space partitioning but not for the transformation subspace training in the subsequent processing. A 2-D example is given in Fig. 3.2(b). It is easily observed that the whole space is more uniformly partitioned, which will bring benefits in local subspace learning.

3.3 Query-driving Appearance Learning and Matching

3.3.1 Local Subspace Learning

By building a tree with multiple levels, the original data points are partitioned into small sets. For each of these sets or clusters k, a spatial transformation A_k is generated. Thus, instead of global dimension reduction as Eq. 2.1, the transformation will become,

$$x_k = A_k F_k = [a_{1,k}^T, a_{2,k}^T, ..., a_{w\times h,k}^T] F_k, a_{j,k}^T \in R^d \tag{3.2}$$

In the subspace modeling, the distance between two points x_1 and x_2, in localized subspace A_k, can be defined as

$$d(x_1, x_2, A_k) = ||A_k(x_1 - x_2)||^2 \tag{3.3}$$

3.3.2 Query-driven Matching Solution

In Chapter 2 Luminance Aligned Projection Distance (LAPD) [105] is proposed to compute the distance between two trajectories. The definition is:

$$d(x, y) = \inf_h \sum_t ||y(t + h) - x(t)||^2 \tag{3.4}$$

where y is the longer trajectory and x the shorter one, and h represents all possible offsets. Since the video clips have different durations, the shorter trajectories are used to match the longer one. For each possible offset, the two trajectories are aligned point by point and the subspace distance between them is calculated. By trying all possible offsets, one optimal distance is finally selected as a measurement of the similarity between the two trajectories.

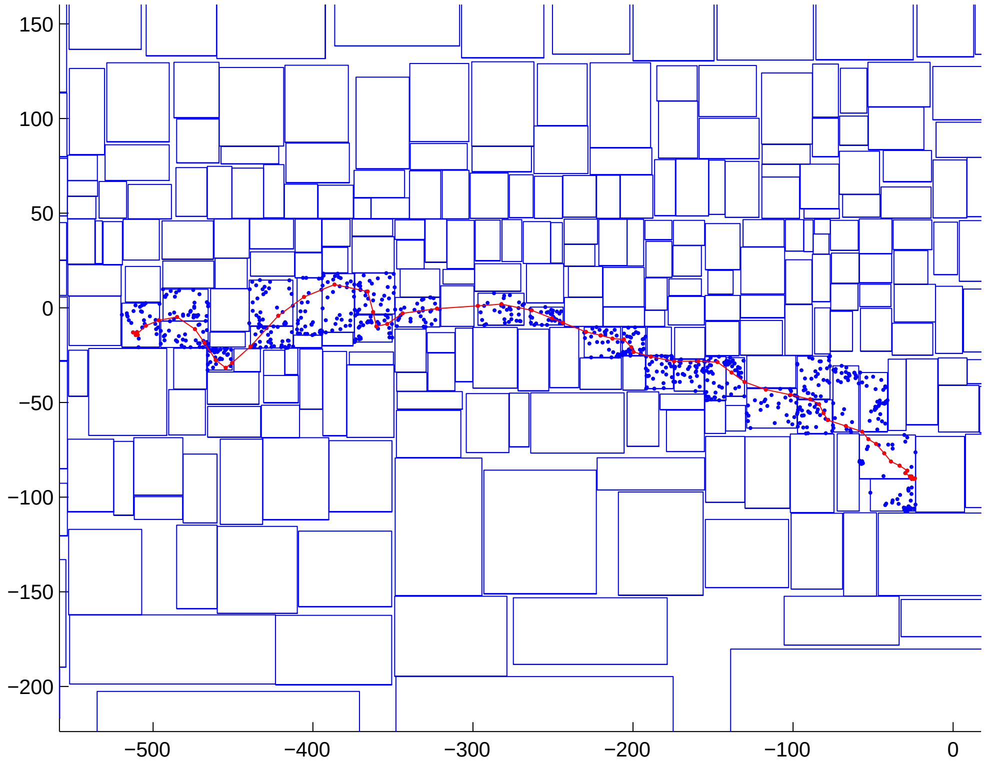


Figure 3.3. An example of Query Trajectory in the Partitioned Space

In this work the reliability of the LAPD is generally improved by adding one more component, the local transformation matrix A_k, into the evaluation of the distance. An example of a query is shown in Fig. 3.3. In the figure the query trajectory is plotted in red, with a background of appearance space partitioned by kd-tree splitting. There is an individual subspace A_k for each leaf node on the tree, which is represented as small rectangular in the figure. When we compute the LAPD between the query trajectory with other trajectories in the training set, we will use corresponding localized subspace to replace the uniform subspace in LAPD solution. Which subspace A_k will be used is determined by where the query clip is positioned as each frame of it traverses the tree. The proposed distance is named as

Localized Luminance Aligned Projection Distance (LLAPD) and defined as,

$$d(x, y, A_k) = \inf_h \sum_t ||A_k(y(t+h) - x(t))||^2 \tag{3.5}$$

Multiple subspaces $A_1, A_2, ..., A_K$ are obtained via PCA after space partitioning. In this step the subspaces are utilized to compute the distance in a localized model. The query trajectory will select a subspace, in which the distance will be computed. In this way, given a query clip $q(t)$, the distance between $q(t)$ and each trajectory in the learning set can be computed and the query label can be determined as the same of the nearest one. That is,

$$j^* = \arg\min_j d(q_k, x_j, A_{q_k}) \tag{3.6}$$

where q_k is the query clip, x_j represents the trajectories in the learning set with label j, d is the distance metric computed by Eq. 3.5 and A_{q_k} is the subspace selected by q_k.

The reliability of similarity measurement is improved by introducing the local appearance space partitioning and indexing. The query clip is also emphasized in the matching solution, compared with the training clip.

3.4 Warping-based Aligned Projection Distance

3.4.1 Dynamic Time Warping

Dynamic Time Warping (DTW) is an algorithm for measuring similarity between time series sequences. It has been widely used in various pattern recognition applications, such as handwriting recognition [3], gesture recognition [13], speech recognition [26], signature recognition [102], shape recognition [54] and others. Time series data is always huge in size, especially when accumulated for a long time in different feature domains. DTW is receiving a lot of interest in data mining and related applications, such as clustering [38] [61], classification [57] [48], association mining [71], and motif discovery [55].

DTW is an extension of Linear Time Warping (LTW), which matches two time series based on a linear alignment of the two temporal dimensions. DTW extend the linear alignment to dynamic programming and search a subspace in which the two sequence have minimum distance between them. Generally, there should be certain constraints imposed on DTW to optimize and expedite the search of the warping path. Major constraints outlined in [69] include monotonic condition, continuity condition, boundary condition, adjustment window condition, and slope constraint condition.

Use our problem as an example, given two video sequences, training clip X and query clip Q as follows: $X = x_1, x_2, ..., x_M$ and $Q = q_1, q_2, ..., q_N$, DTW compute an optimal warping path between X and Q as :

$$D_{DTW}(i,j) = d(x_i, q_j) + \min(D_{DTW}(i-1, j-1), D_{DTW}(i-1, j), D_{DTW}(i, j-1)) \tag{3.7}$$

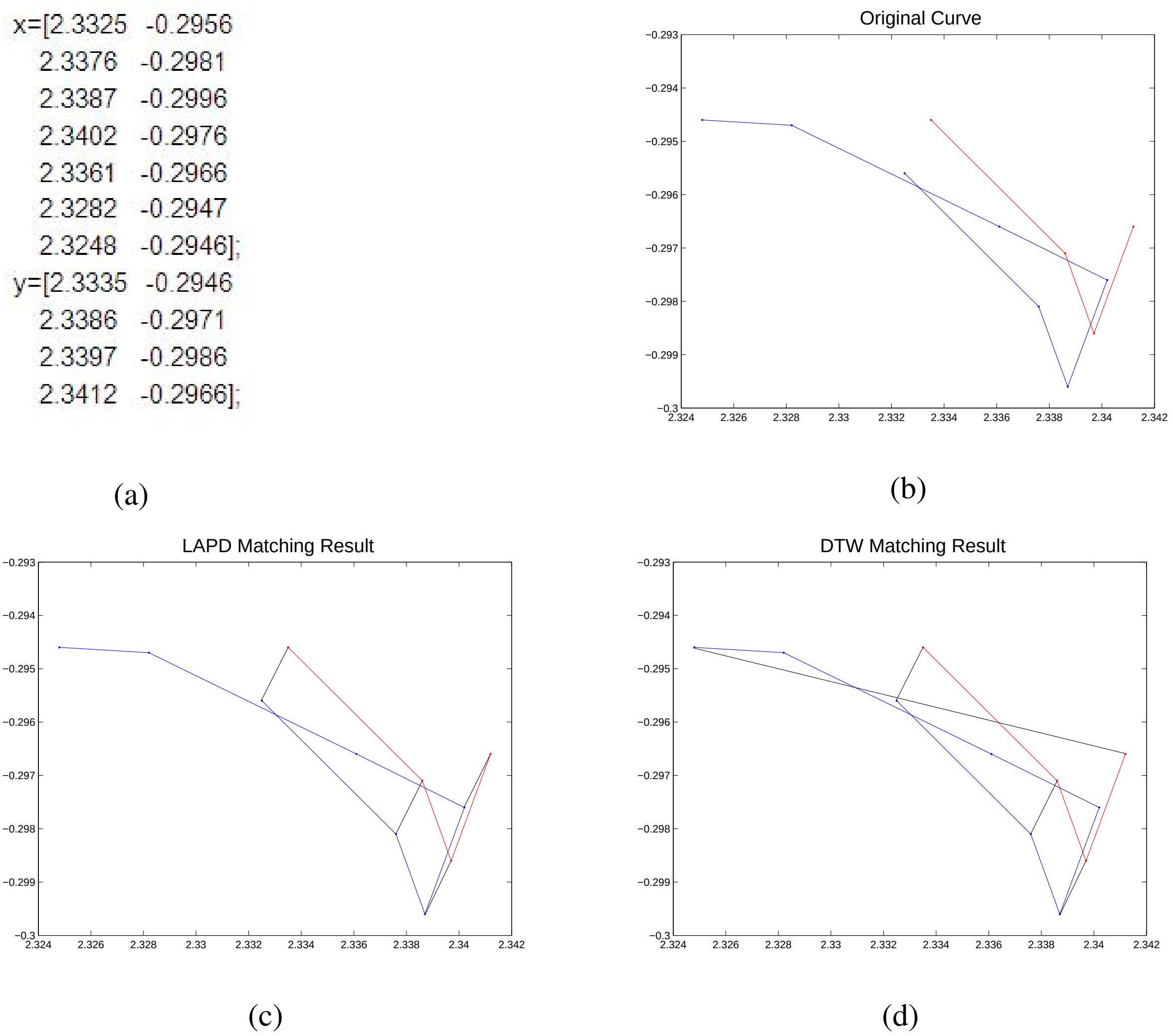
where the $D_{DTW}(i,j)$ refers to the warping distance and $d(x,y)$ stands for the original distance, such as Euclidean one. The computation is completed by using dynamic programming to calculate the minimal cumulative distance recursively.

3.4.2 Warped Alignment in Trajectory Matching

Both global and local appearance modeling focused on improving the description of similarity between trajectories. The aligned projection distance, which is firstly proposed in [105] and improved by local indexing in [106], is widely used in the similarity measurement. This distance is highly dependent on point-to-point distance in the subspace, and the time order is strictly followed. In other words, if kth frame in the query sequence is matched with lth frame in the training sequence, then the $k+1$ query frame is forced to be matched with the $l+1$th frame in the training clip. The objective is to minimize the average point-to-point distance between trajectories, instead of optimizing each matching pair, which should be more reliable.

Motivated by the idea from DTW, in this section we solve the pairwise distance reliability improvement by proposing a warping-based solution. During the alignment processing, flexibility is also considered for a better matching. In the proposed solution, instead of one-to-one matching, a one-to-multiple scenario is considered and a best match will be selected for similarity representation.

On the other hand, the DTW algorithm also has limitations on the trajectory matching task. Each of the point on the trajectory, including some noisy frames, are utilized in the computation of distance. This makes the distance much less reliable as a measurement of similarity. Although DTW offers flexibility in the middle and scaling to both trajectories will not affect the result, this constraint greatly limits the performance.



```
x=[2.3325  -0.2956
   2.3376  -0.2981
   2.3387  -0.2996
   2.3402  -0.2976
   2.3361  -0.2966
   2.3282  -0.2947
   2.3248  -0.2946];
y=[2.3335  -0.2946
   2.3386  -0.2971
   2.3397  -0.2986
   2.3412  -0.2966];
```

(a)

(b)

(c)

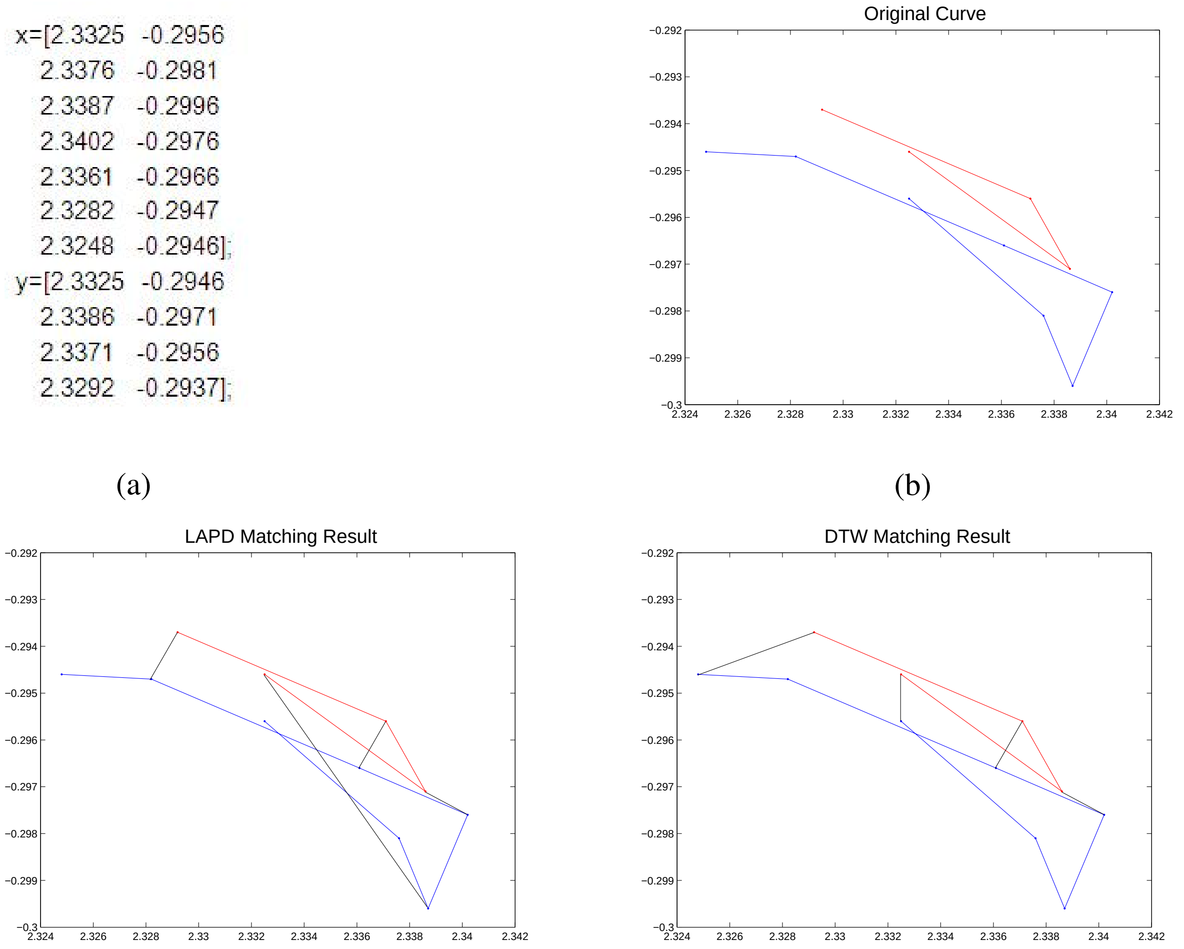
(d)

Figure 3.4. An example for comparison between LAPD and DTW trajectory matching (a) Original data value (b) Original curve (c) LAPD matching result (d) DTW matching result.

To effectively compare the advantage and drawback between LAPD solution and

DTW algorithm, we generate two random example to show the solution respectively.

Example one is presented in Fig. 3.4. We are using a group of 2-D data points to show the trajectory matching. Figure 3.4(a) shows the numerical value of data, and their trajectories are plotted in Fig. 3.4(b). The two algorithms will be applied to match the short trajectory to the long one. Figure 3.4(c) and (d) show the result for LAPD match and DTW match respectively. The matching is plotted as a straight line between the trajectories, from the figure we can see that the LAPD gives a more reliable matching result, while DTW fails to give a good performance due to the first/last points force matching.



```
x=[2.3325  -0.2956
   2.3376  -0.2981
   2.3387  -0.2996
   2.3402  -0.2976
   2.3361  -0.2966
   2.3282  -0.2947
   2.3248  -0.2946];
y=[2.3325  -0.2946
   2.3386  -0.2971
   2.3371  -0.2956
   2.3292  -0.2937];
```

(a) (b)

(c) (d)

Figure 3.5. Another example for comparison between LAPD and DTW trajectory matching (a) Original data value (b) Original curve (c) LAPD matching result (d) DTW matching result.

Another example is shown in Fig. 3.5. The data we used is quite similar as the previous example, only one point on the shorter trajectory is changed. Numerical value is given in Fig. 3.5(a), and trajectory plotted in Fig. 3.5(b). The matching results by LAPD and DTW is shown in Fig. 3.5(c) and (d) separately. Different from the previous example, DTW outperforms LAPD in this case. The reason is that two trajectories have similar appearance, such that degrade the matching performance. Therefore DTW takes advantages from its flexibility in the middle block of the trajectories.

The two examples above are quite special and we are generating a conclusion as follow, if the two trajectories have similar head and tail points but different in length, DTW is a better solution; if they are similar in length but have very different beginning and ending, LAPD should be a better choice. In our problem, most video clips with different action labels have quite obvious different appearance, therefore we emphasized more on LAPD and proposed a Warping-based LAPD solution.

The new proposed solution allows both linear and non-linear point-to-point matching. The appearance similarity is maintained by flexibly trying possible pairwise matching, while the temporal domain constrained such flexibility. If the kth point on one trajectory is matched with lth point on the the other, a cross is defined as matching $k+n$th point is matched with $l-m$th, where m and n are both positive integers. No cross matching is allowed according to temporal constrains, which guarantees the order in time series. With such flexibility, optimal local distance could be found and optimal matching between trajectories are obtained.

3.5 Experiments and Performance Evaluation

In this section we briefly introduce the dataset we use to test our proposed algorithm. Two dataset we used in Chapter 2, Cambridge Hand Gesture Dataset and Youtube Dataset, are also included in this chapter. We also use the Weizmann Dataset to prove the effectiveness of the proposed algorithm. We will present our experiment result and give corresponding analysis.

3.5.1 Recognition Accuracy

3.5.1.1 Weizmann Human Action Dataset

In order to demonstrate the effectiveness of our proposed method, we tested it with the Weizmann action recognition dataset [28]. 10 different actions, $Walk$, Run, $Jump$, $Sideways$, $Bend$, $Wave\ one\ hand$, $Wave\ two\ hands$, $Jump\ in\ place$, $Jump\ in\ jack$ and $skip$, are included in the dataset. Each action is repeated by 9 different persons and 90 video clip are involved in total. Example frames are shown in Fig. 3.6.

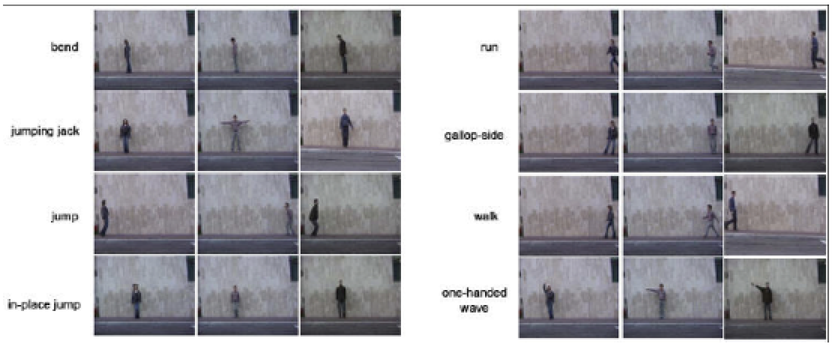
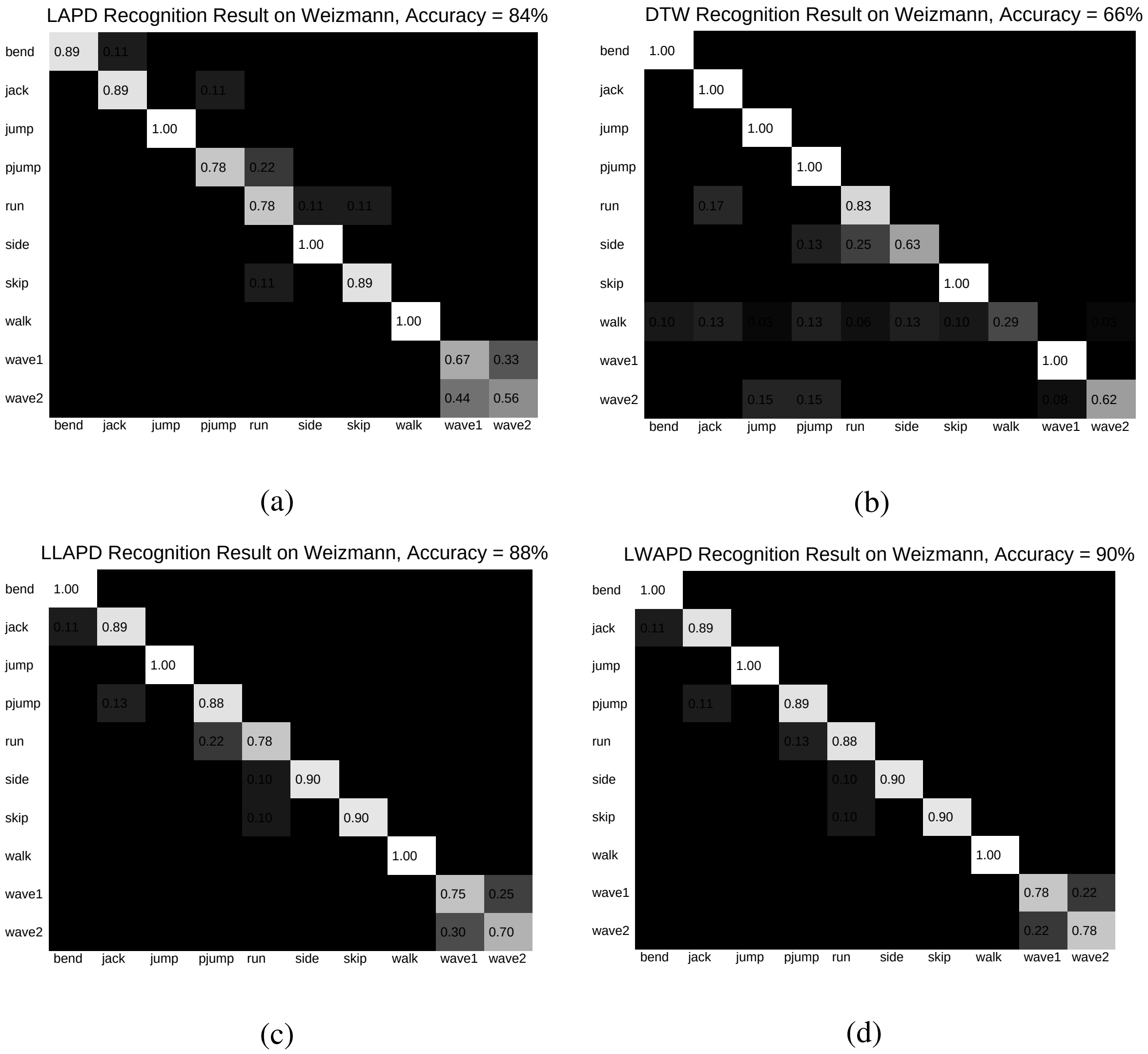


Figure 3.6. Sample frames of Weizmann Dataset

In the preprocessing, the clips are first down-sampled to 32 by 24 icons and then a global PCA is applied to further reduce the number of dimensionality down to 64. 7-level indexing is carried out on all frames and localized subspaces are trained for each sub-block.

We present result on this dataset by 4 different scenarios: Luminance Aligned Projection Distance (LAPD) approach, Dynamic Time Warping (DTW) approach, Localized Luminance Aligned Projection Distance (LLAPD) and Localized Warpped Aligned Projection Distance (LWAPD) approach. Different similarity measurement schemes are applied to compute the distance between each pair of clips, and all clips are tested with a "leave one actor out" scenario. The simulation results are shown in Fig. 3.7.

Result of LAPD is shown in Fig. 3.7(a), a recognition accuracy of 84% is achieved by this approach. Figure 3.7(b) present the result obtained by DTW, with an accuracy of 65%.



LAPD Recognition Result on Weizmann, Accuracy = 84%

(a)

DTW Recognition Result on Weizmann, Accuracy = 66%

(b)

LLAPD Recognition Result on Weizmann, Accuracy = 88%

(c)

LWAPD Recognition Result on Weizmann, Accuracy = 90%

(d)

Figure 3.7. Recognition accuracy on Weizmann Action Dataset, (a) LAPD approach, (b) DTW approach, (c) LLAPD approach, (d) LWAPD approach.

These two scenarios are used as a baseline for comparison of our proposed schemes. In Fig. 3.7(c) we plot the confusion matrix of LLAPD recognition result. Hierarchical indexing improved the result to 88%. The numerical result is improved to 90% by LWAPD, as plotted in Fig. 3.7(d).

It can be observed from the confusion map that the two types of waving are confused in several cases and this maybe due to loss of detail information during pre-processing. The overall recognition result from LLAPD is 88%, which is comparable with the result in the literature without manual labeling or cropping. The local warping offers more flexibility for point-to-point trajectory matching and further improve the result by 2 percentage.

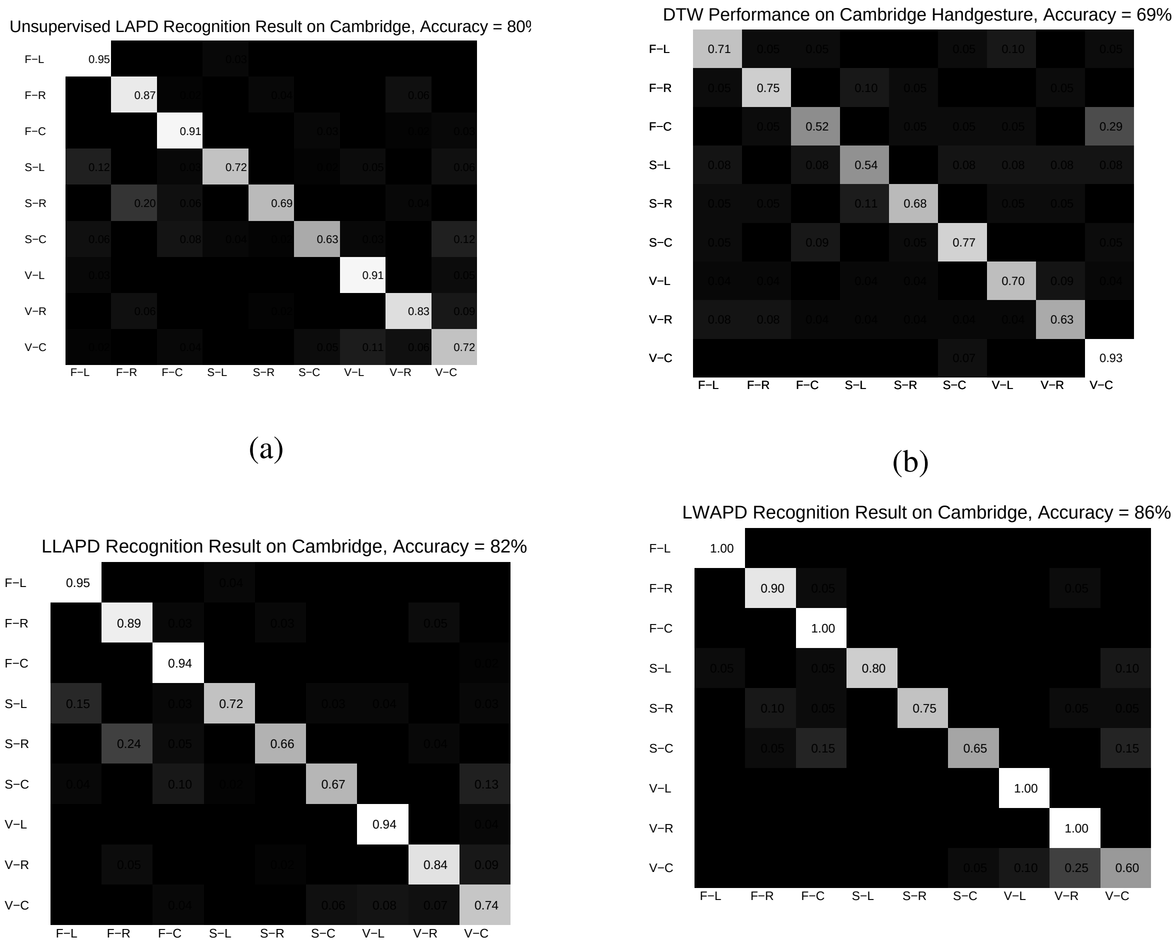
From the figure we can generally see the LAPD is discriminative over different action classes. Several mistake is uniformly distributed in different actions. DTW is not very good at such kinds of classification, as proved by experiment result in Chapter 2. The main reason is the ”begin and end curse”, which is a basic assumption in DTW solution. However, in practice, the starting and ending frame of video sequences with same action label may have quite different appearance, and therefore greatly degrade the performance. On the other side, LLAPD outperforms LAPD by several percent, and further improved by L-WAPD. It is demonstrated that local information and flexibility in matching will be helpful to better evaluate similarities between different high-dimensional trajectories, even though the improvement is not very much.

3.5.1.2 Cambridge Hand Gesture Dataset

The detail information can be found in Chapter 2, so here we only list several critical numbers. This dataset is composed by 900 image sequences. 9 different action labels are included by combining 3 primitive hand shape, *Flat (F)*, *Spread (S)* and *V-shape (V)*, and 3 primitive motion directions, *Leftward (L)*, *Rightward (R)* and *Contract (C)*. 20 subjects are captured in 5 different luminance condition, for each action class [42].

We are using the “Leave one subject out” strategy to test our proposed algorithm. Noted for LAPD, we are using the unsupervised version for a fair comparison. The proposed indexing approach utilized the local information but failed to form a supervised version since the label information is missing for each subblock. The baseline performance is shown in Fig. 2.12(b) and (d) respectively for LAPD and DTW. We repeat the figure here as Fig. 3.8(a) and (b) for comparison convenience. Numerical recognition accuracy for LAPD and DTW are 80% and 69% respectively. Similar as the Weizmann dataset result, DTW is not performing good due to the inaccuracy in similarity measurement.

Numerical result is improved to 82% by LLAPD approach, as shown in Fig. 3.8(c). In Fig. 3.8(d) warping further improved this number by 4 percentage to 86%. Theoretically,



(a)

(b)

(c)

(d)

Figure 3.8. Recognition accuracy on Cambridge Hand Gesture Dataset, (a) LAPD approach, (b) DTW approach, (c) LLAPD approach, (d) LWAPD approach.

the LLAPD takes care about the local statistical appearance relationship, while the LWAPD improves more flexibility on temporal domain. Intuitively the motion in this dataset is slight, so the temporal flexibility is much more important, which also bring larger improvement in recognition accuracy.

3.5.1.3 Youtube Dataset

Youtube Dataset is tested in Chapter 2, for our proposed DLFT and LAPD approach. This set is composed of 1,577 outdoor sports video sequence [50]. Totally 11 different sports actions

Table 3.1. Hand Gesture Recognition Accuracy Comparison(%).

Method	Set1	Set2	Set3	Set4	Set5	Average
DLFT(Unsupervised)	76	75	68	74	72	73.0
LAPD(Unsupervised)	83	81	77	80	79	80.0
LLAPD	84	83	81	84	79	82.2
LWAPD	89	85	83	89	84	86.1
TCCA [42]	81	81	78	86	–	81.5
Nieble [58]	70	57	68	71	–	66

are included. 25 persons repeated a single action for a few times. This set is challenging due to the complicated background, which is not contributive to the human action recognition.

In the pre-processing step we resample the original video sequence down to 40 by 32 icons. Global PCA is applied to the reduce the number of dimension down to 64. Trajectories in 64-dimensional space are used as a representation of the video clips.

The dataset is composed by more than 250,000 frames, so we apply a 9-level appearance space partitioning. 10,000 white noise points are added in order to make the partitioning more reliable. Under the K-d tree structure, there are totally 512 leaf nodes, each of which contains about 500 data points.

We present the experiment result in Fig. 3.9 for 4 different scenarios. Unsupervised LAPD is giving an accuracy of 91% as shown in Fig. 3.9(a), which already outperforms other techniques in the literature [51]. The result of DTW is about 69%. The very different starting and ending frame appearance greatly degrade the performance of traditional DTW approach.

Our proposed solution LLAPD improved the recognition result by only 1 percent as shown in Fig. 3.9(c). The local information is slightly helpful in constructing a more reliable distance metric in the dataset. Due to the good result achieved by global subspace learning in LAPD, the room for improvement is limited and there is not much numerical enhancement.

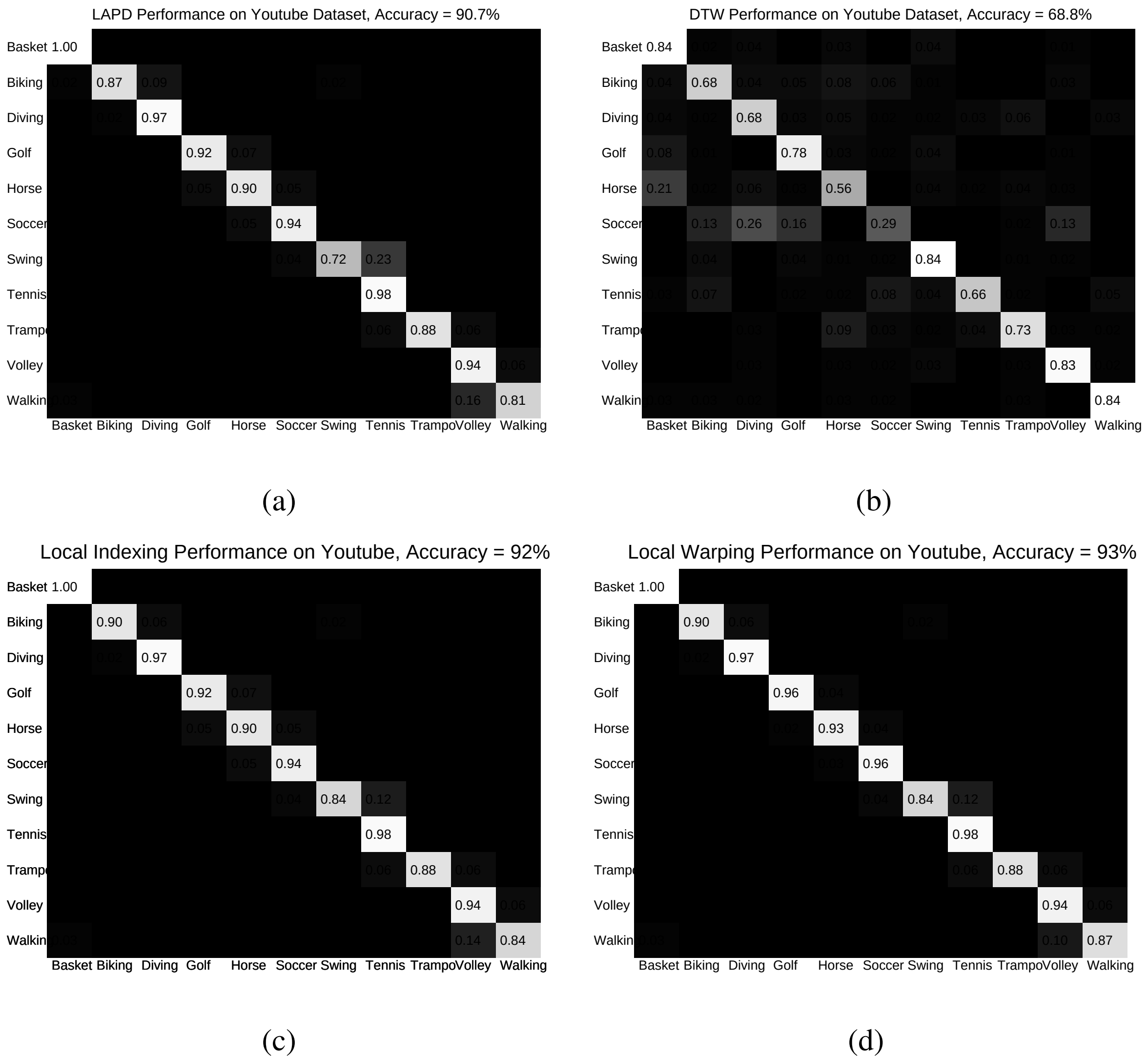


Figure 3.9. Recognition accuracy on Youtube Dataset, (a) LAPD approach, (b) DTW approach, (c) LLAPD approach, (d) LWAPD approach.

Result of LWAPD is plotted in Fig. 3.9(d), and compared with LLAPD, the contribution brought by warping is also not very much, also about 1 percent for recognition accuracy.

Both of the LLAPD and LWAPD are a non-degradation solution of global LAPD. The similarity between different video trajectories are better described and more reliable distance are computed during the processing. LAPD is a special case of LLAPD when all the local subspace are uniformed as unit matrix, i.e., the distance is computed in Euclidean way. Also, LAPD is a special case of LWAPD when there is no flexibility allowed, i.e., all the points should be matching one by one without any overlapping.

Table 3.2. Computational Complexity

	$L = 8$	$L = 10$	$L = 11$
Locality Search	$133.2ms$	$115.6ms$	$91.0ms$
Local PCA	$201.6ms$	$94.8ms$	$66.0ms$

3.5.2 Time Complexity

In this section we present the timing issue for our proposed algorithm. In most applications there is always a tradeoff between performance and complexity. The better performance in accuracy is obtained at a cost of query-driven local modeling. The computational complexity involved are appearance space partitioning and indexing, which is offline, as well as the locality searching and local model building from queries, which is online. The online complexity part is summarized in Table 4.2, for the time spent on locality searching and local PCA on a 2.0 Ghz laptop PC running Matlab, for different granularity of appearance space partitioning as indexing tree height L.

When the number of level, L, is larger, a longer time is needed for partitioning and indexing. Meanwhile, as shown in Table 4.2, the online steps, training local is less time-consuming when L is larger. This is due to a smaller number of points in each partitioning space, which make the computation for both local searching and local PCA faster. Besides, the selection of L is a tradeoff in accuracy. Too large of an L may result in overfitting problem while a smaller L cannot offer accurate modeling.

3.6 Summary

In this chapter we developed two new solutions for video action recognition and hand gesture recognition based on the Luminance Aligned Projection Distance. The appearance space is partitioned and indexed for better model localization and fast access. Temporal flexibility is considered and utilized to provide better matches.

Query driven local appearance models are invoked at query time for distance metric computing during the video query trajectory matching. The solutions addressed the appearance complexity issue effectively and efficiently. Warping-based LAPD provides more temporal flexibilities on trajectory matching. Simulation results on three different datasets demonstrated its effectiveness in recognition accuracy as compared to the state of the art, and robustness as well. The overall computational efficiency is also good, due to the efficient appearance space indexing. For applications with large set of action labels, this framework is especially effective and can achieve satisfactory performance in real time. The framework can also be mapped to other temporal matching schemes like HMM and DBN for more complex temporal behavior matching.

CHAPTER 4

MULTI-MODALITY DISTANCE METRIC CO-LEARNING FOR AUDIO-VISUAL SPEAKER IDENTIFICATIONS

4.1 Overview

Within the last several years, there were numerous algorithms about statistical pattern recognition proposed to apply for face detection, object tracking and other video pattern recognition problems. On the other hand, the advances in computing and communication technologies make the capture, communication and storage of video content much easier and cheaper than ever. The demand of visual content analysis is growing rapidly and there are many related applications.

To enable more intelligent and meaningful applications with this vast amount of video data, especially in surveillance and on-line video repository searching and mining, a fast and effective video content analysis and learning solution is becoming a crucial task in the effort of making video content more accessible and intelligent. Example applications include surveillance video analysis for security, sports video analysis for labeling and searching.

General applications such as video face recognition, landmark identification and action/gesture classification are challenging due to the processing of large amount of data. With the development of capture devices, millions of pixel values are included in a single video frame, and there are also hundreds of frames per minute. In this way, a video sequence with several minutes duration is composed by billions of pixel values, which brought difficulties for fast analysis and understanding. However, what we need for understanding is rather simple: just find out who the person is, whether there is a pre-defined event or not, what kind of action/gesture is performed and so on. Therefore it is generally a complicated procedure to process the video, and various kinds of algorithms to simplify such processing are proposed in the literature.

Traditional video pattern recognition approaches focused on the motion estimation and object segmentations, especially object tracking. These approaches are not practical when there is no distinct pre-definition for object and will not work effectively when the environment is changing such as lighting, background conditions, and objects. Vast amount of research work exists in this area. Typical approaches include object segmentation and motion estimation and based solutions, [18] [100] [103], which try to explicitly recover object and motion information and then recognize human action based on learning of object level spatio-temporal primitives. Such approach suffers from poor robustness to the appearance variances of human actions with different subjects appearance, lighting, background conditions and occlusions.

In this chapter we focus on solving a special video pattern recognition problem, speaker identification, which means identifying speakers from video sequence, together with the audio information. In the machine learning literature, the speaker recognition problem can be categorized into two problems: speaker identification and speaker verification [14]. Verification is a one on one matching problem to justify whether the query speaker is the same person as the labeled speaker. This is very similar to a true or false question. Speaker identification, which is more complicated and challenging than verification, is a task of labeling unknown voice to one of a set of speakers.

Extensive research efforts have been dedicated to the acoustic speaker identification. Given the audio signals, source-filter models are widely applied, which leads to the extraction of features such as Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP) and Mel-Frequency Cepstrum Coefficients (MFCCs) [32]. In the traditional approaches, models are constructed based on a set of training data and then used for the identification task. However, speaker identification based on pure audio information failed to achieve high accuracy in many practical applications, especially when there are noises in background or a bias exists between training data and testing data. On the other hand, visual information is very descriptive for the appearances of speakers. The visual features contain more discriminative information and greatly improve the accuracy in the speaker identification. Therefore, several research works focused on investigating the audio and visual information simultaneously to process the identification task, which is the so-called audio-visual speaker identification [2] [52].

In multimedia content understanding and retrieval literature, the simultaneous investigation of information from multiple channels is usually named multi-modality analysis. In the literature the multi-modality analysis is extensively studied, there are two typical approaches [78] [89]. One approach is to connect the features extracted from different modalities and then learn with the joined features. Another approach is to learn with each modality individually and then fuse the results, such as [37]. However, these two approaches do not take the relationship between the modalities into account, which actually can benefit the learning process. In machine learning literature, several algorithms are proposed to deal with such problems, which is named multi-view learning [59] [74] [75]. These methods are built based on the consistency of different modalities. More specifically, it is assumed that the classifiers learned from the two views should predict closely on unlabeled data. Despite these methods are able to achieve better performance than the fusion methods, they require a large amount of unlabeled data, and limits such algorithms to many applications.

In this chapter we propose a multi-modality distance metric co-learning approach. It can be applied for various tasks, including audio-visual speaker identification, text analysis and digit recognition. Compared with previous solutions in the literature, our approach only needs labeled data during the training, which is practical in our real world. The structural consistency between two modalities is utilized to learn a representative distance metrics, by which the performance can be improved.

Figure 4.1 demonstrates a schematic illustration of our approach. Firstly we have audio and visual features respectively and then we put them together to generate a group of comprehensive feature, the so-called multi-view features. Distance metric learning and classifier modeling are applied on these features and classifiers are obtained via modeling. Finally, given a query sample for testing, the identification result is determined by the classifier.

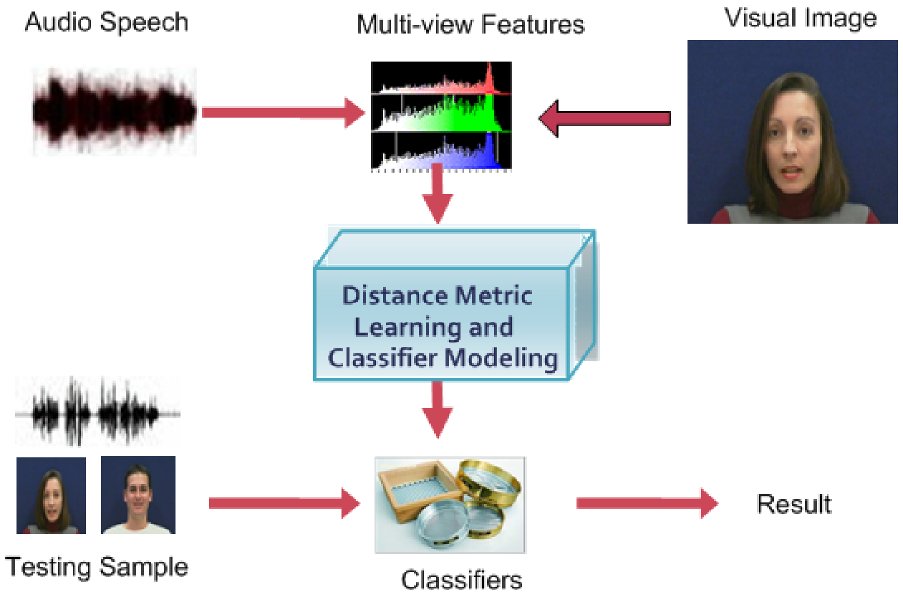


Figure 4.1. Audio and Visual Modalities in proposed approach

The remaining of this chapter is organized as follows. In section 4.2 we briefly review the techniques for machine learning problems which are widely applied. In section 4.3 we propose our multi-modality algorithm followed by a distance metric learning approach. The experiment setup and result are presented in section 4.4 and also analyzed. Finally we conclude our work in section 4.5 and propose some future ideas.

4.2 Related Work

Multi-modality learning is proposed to deal with problems that involve multiple set of features. In our speaker identification problem, speakers can be recognized from the acoustic signal and the visual appearance. Considering in practise, there may not be enough labeled training data, most multi-modality learning methods use a large amount unlabeled data, which is called semi-supervised learning (SSL) [74] [75]. It is assumed that the unlabeled data have similar distribution with the labeled ones [7] [10] [59]. However, this assumption is not always true in real cases. Besides, in many applications the requirement of large unlabeled set may be infeasible. To take advantage of the unlabeled data with a reliable processing, Canonical Correlation Analysis (CCA) and kernel CCA [29] adopt a different approach. Features from different modalities are directly extracted and classifiers are trained based on the consistency between modalities. The label information of training data is not used either in feature extraction or learning, so it is not promised to have a discriminative performance. Multiple kernel models are constructed in [94] for image annotation, which is a multiple label prediction task.

Distance metric learning targets on constructing an optimal metric for the given learning task based on the pairwise relationships among samples. A number of algorithms have been proposed for distance metric learning [99]. Relevant Components Analysis (RCA) method is proposed to learn a linear transformation from equivalence constrains, which can be used directly to pairwise compute the distance [34]. In [92] distance metric learning is formulated as constrained convex programming by minimizing the distance between the data points in the same classes. The data sample from different classes are clearly separated. Neighborhood Component Analysis (NCA) [27] learned a distance metric by extending the nearest neighbor classifier. Large margin nearest neighbor (LMNN) method is presented by Weinberger in [87], the authors extended NCA through a maximum margin framework.In order to better present our method.

In the following section we will review several classical algorithms in the feature extraction. The first one is the Principle Component Analysis (PCA), which is firstly proposed in [84] as a global subspace learning method, to processing features with dimensionality reduction. PCA targets on finding the low-dimensional embedding of data points that best preserves the variance of data set. Linear Discriminate Analysis (LDA), proposed in [4] and applied on face recognition is a supervised method to extract information. Both PCA and LDA are linear subspace learning method. As a non-linear solution, the Support Vector Machine (SVM) is proposed [12] [97].

4.2.1 Principle Component Analysis

PCA is a useful statistic technique that has found application in face recognition and related pattern recognition applications. Usually in multimedia processing, it is easy to find features in a very high dimension. Due to the redundancy in video, the high dimensionality of features is not helpful for classification, but cost a lot of unnecessary computation resource. Therefore, the dimensionality reduction is one of the critical processing steps.

Mathematically, given n data points $x_1, x_2, ..., x_n \in R^D$, PCA aims to find the optimal subspace A on which the following objective function is maximized.

$$F = \arg \max_{A^T A = I^d} tr(A^T C A) \tag{4.1}$$

where I_d is the $d \times d$ identity matrix, $tr(.)$ refers to the trace operator, and C is defined as follow,

$$C = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \tag{4.2}$$

where $\bar{x}$ denotes the mean value of all x data. In order to solve the optimization problem, the following computational steps are presented in the next paragraph.

PCA is composed by the following several steps to extract the patterns in a group of data in high dimension space. Considering we have n data, each with a dimension of D, we can put these data into a matrix, S, with size n by D. Firstly the mean value is calculated and subtracted from the original raw data, then the covariance matrix for the obtained matrix is calculated, which has a size of D by D. Then the eigenvalues and eigenvectors of the covariance matrix are generated, and form another D by D matrix with one single eigenvector in each column. Smaller eigenvalue means the corresponding eigenvector is not as informative as the larger ones. By discarding some of these columns of less importance, the number of dimensions is reduced. On the other hand, by preserving the most important eigenvectors, it means the correlation of the dimension is removed and the data are projected into a new space, whose bases are orthogonal with each other. The computational step is plotted in Fig. 4.2.

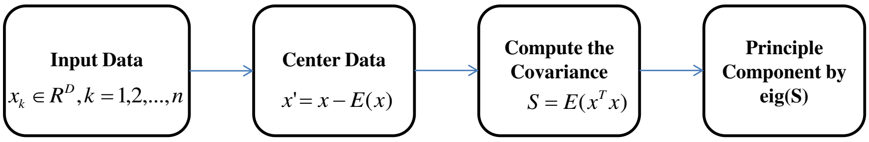


Figure 4.2. PCA Computation Steps

PCA complete the dimensionality reduction by preserving the most important statistical information. Regarding the redundancy in the video, the lossy caused by PCA will not degrade the classification result. PCA is an unsupervised global subspace learning, which did not utilize the label information for subspace learning and discrimination, which affects the performance on reliability. In order to have better usage of the training data samples, LDA is proposed as a supervised learning method.

4.2.2 Linear Discriminant Analysis

LDA is another subspace learning method with dimensionality reduction. It is different from PCA because of it goes from descriptive level to a discriminate level. Instead of collecting data blindly, the LDA used label to do a supervised learning to offer a boundary which is used to distinguish and do classification. This also brought another several problems such as how many features need to be extracted and how to define the discriminant and separability between classes. The computational steps involved is shown in Fig. 4.3.

Different from PCA, LDA maps the high dimensional data points to a lower dimensional space by minimizing the with-in class variance and maximizing the between-class variance simultaneously. By finding such a subspace A, the objective function is maximized as follow,

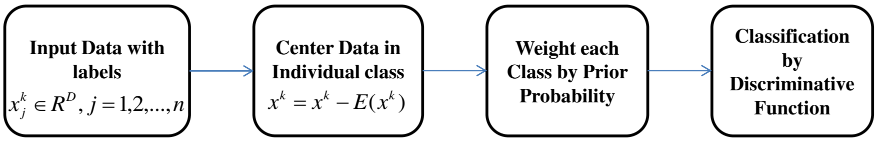

Input Data with labels
$x_j^k \in R^D, j = 1,2,...,n$
Center Data in Individual class
$x^k = x^k - E(x^k)$
Weight each Class by Prior Probability
Classification by Discriminative Function

Figure 4.3. LDA Computation Steps

$$F = \arg\max \frac{tr(A^T S_b A)}{tr(A^T S_w A)} \tag{4.3}$$

where the S_b and S_w are defined as the between-class variance and with-in class variance as follow respectively.

$$\begin{aligned} S_b &= \sum_{c=1}^{r} n_c(\bar{x_c} - \bar{x})(\bar{x_c} - \bar{x})^T \\ S_w &= \sum_{c=1}^{r} \sum_{j \in \Omega_c} (x_j - \bar{x_c})(x_j - \bar{x_c})^T \end{aligned} \tag{4.4}$$

where the r is number of classes in the problem, n_c and $\bar{x_c}$ are the number of data samples belongs to the cth class and their mean value separately. The Ω_c is the set of data samples which belongs to the cth class. The solution of the optimization problem is solved by generalized eigenvalue decomposition as follow,

$$S_b a_i = \lambda_i S_w a_i \tag{4.5}$$

where a_i is a generalized eigenvector.

In summary, PCA is an unsupervised global method, and is usually using to find a subspace that put as much data in the set as possible. LDA is a supervised method which depends on the training data, and mainly works on scattering the data in the different classes. Figure 4.4 shows the comparison between these two methods.

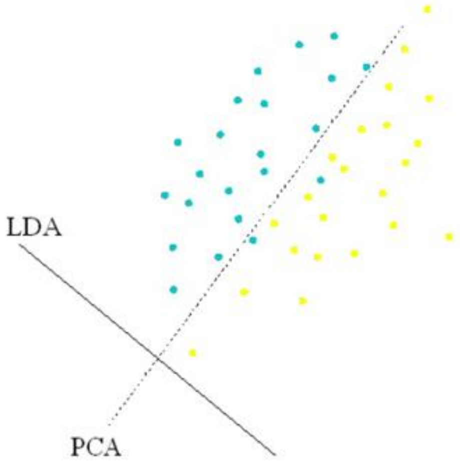


Figure 4.4. Difference between PCA and LDA for subspace learning: Unsupervised VS. Supervised

4.2.3 Support Vector Machine

Support Vector Machine is firstly proposed in [12] for pattern recognition. SVM aims on minimizing the structural risk rather than the empirical risk, to avoid the problem brought by over fitting.

SVM convert the classification problem into a constrained optimization problem by trying to maximize the gap between different classes to generate a classifier with good performance. The key approach for a constrained optimization is using the Lagrangian Multiplier, during the process of optimization, some of the factors contribute to classification while others not. These “useful” factors are named as the “support vector”, which are highly related to determine the boundary. Classifier can be designed only with the useful factors and the factors with less importance can be discarded.

One example is given in the Fig. 4.5. Particularly, SVM can only solve problems with 2 classes but cannot handle more. For multiple classes, usually we use LDA to handle, or a bank of SVM classifier.

For most of the above transformations, there is a kernel version for each of them,

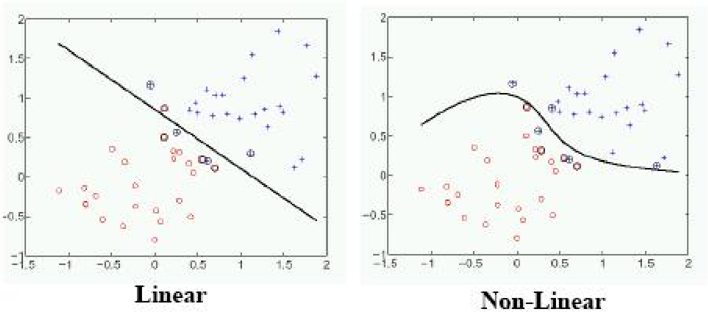


Figure 4.5. SVM Classifier Example: Linear VS. Non-linear

such as kernel PCA, kernel LDA and kernel SVM [97]. The difference is that kernel functions operate in the features space without calculating the coordinates, but only concentrate on computing the inner products between data pairs, which also contains the feature information. Compared with the original version, kernel methods also mapped the data into a higher dimensional feature space with non-linear computations. As shown in Fig. 4.5, the non-linear method provide a more reliable boundary in many cases.

4.2.4 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is firstly proposed in 1936 by Hoteling [36]. It is a method of correlating linear relationships between two multidimensional variables. CCA utilize two modalities of the same semantic object to extract the representation of semantics. It is assumed that two modalities are highly correlated with each other so the correlation can be utilized to learn a linear subspace.

In real applications, there are usually multiple set of features extractable. For example, in image processing, color moment and SIFT features [53] are two groups of different features. In our audio-visual speaker identification problem, audio features are visual features are extracted respectively. Intuitively, the mouth shape of the speaker should varies according to the words pronounced by the speakers. Therefore there should be some correlations between the audio and visual features. CCA takes advantage of thus correlation to learn a group of subspaces.

Consider a multivariate random vector of the form (x, y), suppose we have a sample of instances $S = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$ of (x, y). Denote S_x as $(x_1, x_2, ..., x_n)$ and similarly S_y as $(y_1, y_2, ..., y_n)$. The CCA processing can be considered as finding a new coordinate for x by choosing a new direction $\boldsymbol{w}_x$ and projecting x onto this direction, i.e.,

$$x \rightarrow < \boldsymbol{w}_x, x > \tag{4.6}$$

Correspondingly a direction can be obtained for variable y as $\boldsymbol{w}_y$. The sample of data can be represented with the new coordinate as

$$\begin{aligned} S_{x,\boldsymbol{w}_x} &= (< \boldsymbol{w}_x, x_1 >, < \boldsymbol{w}_x, x_2 >, ... < \boldsymbol{w}_x, x_n >) \\ S_{y,\boldsymbol{w}_y} &= (< \boldsymbol{w}_y, y_1 >, < \boldsymbol{w}_y, y_2 >, ... < \boldsymbol{w}_y, y_n >) \end{aligned} \tag{4.7}$$

In the subspace $(\boldsymbol{w}_x, \boldsymbol{w}_y)$, the correlation between the two set of vectors is maximized. Therefore, the function's result can be equivalently considered as the following problem,

$$\begin{aligned} \rho &= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} corr(S_x \boldsymbol{w}_x, S_y \boldsymbol{w}_y) \\ &= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{< (S_x \boldsymbol{w}_x, S_y \boldsymbol{w}_y >}{||S_x \boldsymbol{w}_x|| ||S_y \boldsymbol{w}_y||} \end{aligned} \tag{4.8}$$

where the $corr(.)$ refers to the correlation computation.

By introducing the $E(f(x, y))$ to denote the empirical expectation of $f(x, y)$, the equation 4.8 can be rewritten as

$$\begin{aligned} \rho &= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{E(< \boldsymbol{w}_x, x >< \boldsymbol{w}_y, y >)}{\sqrt{E(< \boldsymbol{w}_x, x >^2) E(< \boldsymbol{w}_y, y >^2)}} \\ &= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{E(\boldsymbol{w}_x' x y' \boldsymbol{w}_y)}{\sqrt{E(\boldsymbol{w}_x' x x' \boldsymbol{w}_x) E(\boldsymbol{w}_y' y y' \boldsymbol{w}_y)}} \\ &= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x' E(x y') \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x' E(x x') \boldsymbol{w}_x \boldsymbol{w}_y' E(y y') \boldsymbol{w}_y}} \end{aligned} \tag{4.9}$$

The equation 4.9 can be further simplified by introducing the covariance matrix. Denote the within-sets covariance as C_{xx} and C_{yy} and the inter-sets covariance matrix as C_{xy} and C_{yx}. Since we have $C_{xy} = C'_{yx}$. Then we rewrite the function ρ as

$$\rho = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}'_x C_{xy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}'_x C_{xx} w_x w'_y C_{yy} w_y}} \tag{4.10}$$

Therefore, the maximum correlation is the maximum of ρ with respect to $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$. And this can be equivalent to a constraint optimization problem,

$$\begin{aligned} \text{Maximize} \quad & \boldsymbol{w}'_x C_{xy} \boldsymbol{w}_y \\ \text{subject to} \quad & \boldsymbol{w}'_x C_{xx} \boldsymbol{w}_x = 1, \boldsymbol{w}'_y C_{yy} \boldsymbol{w}_y = 1 \end{aligned} \tag{4.11}$$

The optimization problem can be solved by Lagrangian Multiplier, the corresponding Lagrangian function is

$$L(\lambda, \boldsymbol{w}_x, \boldsymbol{w}_y) = \boldsymbol{w}'_x C_{xy} \boldsymbol{w}_y - \frac{\lambda_x}{2}(\boldsymbol{w}'_x C_{xx} \boldsymbol{w}_x - 1) - \frac{\lambda_y}{2}(\boldsymbol{w}'_y C_{yy} \boldsymbol{w}_y - 1) \tag{4.12}$$

The solution can be obtained by taking the partial derivatives in respect to $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ and set them to zero, then we have,

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{w}_x} &= C_{xy} \boldsymbol{w}_y - \lambda_x C_{xx} \boldsymbol{w}_x = 0 \\ \frac{\partial L}{\partial \boldsymbol{w}_y} &= C_{yx} \boldsymbol{w}_x - \lambda_y C_{yy} \boldsymbol{w}_y = 0 \end{aligned} \tag{4.13}$$

Considering in most general case, we assume $\lambda_x = \lambda_y$. Given the equation group above, assume that the covariance matrix is invertible, reshape the 4.13, we can generate the relationship by linear operation.

$$\boldsymbol{w}_y = \frac{C_{xy}^{-1} C_{yx} \boldsymbol{w}_x}{\lambda} \tag{4.14}$$

By substituting Equation 4.14 into Equation 4.13, we have

$$C_{xy} C_{yy}^{-1} C_{yx} \boldsymbol{w}_x = \lambda^2 C_{xx} \boldsymbol{w}_x \tag{4.15}$$

Finally it become a generalized eigenvalue decomposition problem, which is solvable by Cholesky decomposition [29].

Generally, the CCA can be summarized as following steps, firstly compute and decompose the covariance matrix, then find out the corresponding eigenvalue and eigenvector, and finally compute the subspace. The computation is shown in Fig. 4.6.

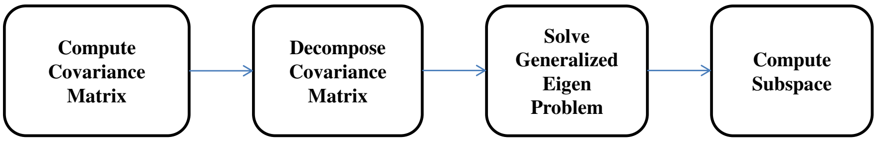

Compute Covariance Matrix
Decompose Covariance Matrix
Solve Generalized Eigen Problem
Compute Subspace

Figure 4.6. CCA Computation Steps

4.2.5 Neighbor Component Analysis

From CCA we borrow the idea on utilizing correlation between different set of features, and our proposed approach is also closely related with the Neighbor Component Analysis (NCA) [27], but it is worth mentioning that our approach is flexible and we can also extend other distance metric learning methods such as LMNN [78]. In this section we firstly brief review the mathematical processing in NCA.

The NCA method is proposed to optimize a distance metric M which minimize the leave-one-out classification error rate of k-NN on training samples. Traditionally, in any subspace $\mathbf{A}$, the distance between x_i and x_j is denoted as $d(x_i, x_j) = (x_i - x_j)^T M(x_i - x_j) = (x_i - x_j)^T A^T A(x_i - x_j)$, specially the distance is Euclidean when $\mathbf{A}$ is unit matrix. This definition of distance is effective, but may not be accurate as a measurement of similarity between two samples.

To solve such problem of k-NN classification error, NCA applied a stochastic neighbor assignment. Particularly, for each point i, a neighborhood j can be selected with a probability p_{ij}, defined as:

$$p_{ij} = \frac{exp(-||A(x_i - x_j)||^2)}{\sum\limits_{k \neq i} exp(-||A(x_i - x_k)||^2)} = \frac{exp(-||A(x_{ij})||^2)}{\sum\limits_{k \neq i} exp(-||A(x_{ik})||^2)} \tag{4.16}$$

From now on, for simplicity we are using x_{ij} to denote $x_i - x_j$. In NCA approach, the optimal solution is obtained by maximizing the expectation of correctly classified sample numbers as follow,

$$\max_A f(A) = \sum_i \sum_j p_{ij} = \sum_i \sum_j \frac{exp(-||A(x_{ij})||^2)}{\sum\limits_{k \neq i} exp(-||A(x_{ik})||^2)} \tag{4.17}$$

It is easily demonstrated that the objective function is differentiable to $\mathbf{A}$, and the partial derivative is as follow:

$$\frac{\partial f}{\partial A} = 2A \sum_i (p_i \sum_k p_{ik} x_{ik}^T x_{ik} - \sum_{j \in C_i} p_{ij} x_{ij}^T x_{ij}) \tag{4.18}$$

Therefore, the problem can be solved by a gradient descent approach and the global optimal can be achieved when the partial derivatives converge.

4.3 Multi-modality Distance Metric Co-learning

We extend traditional distance metric learning to deal with multi-modality problems. The most straightforward approach is to directly learn a distance metric for each set of features and fuse the recognition result. However, it neglects the structural consistency of different modalities and in this section we propose a method to explore this information to learn more reliable distance metric.

We are referring the inter-modality information as feature structure consistency between different modalities. Assuming that we have two group of features, intuitively the distance metrics for different set should have consistency in some way. More specifically, if two samples are closed with each other in one set of feature, the distance between them in the other feature space should not be far away either. On the contrary, if two samples are very faraway from each other in one modality, they are less likely to be close in the other feature space. This is the so-called structural consistency. The distance metric will become more reliable after we take the structural consistency into account.

Based on the consideration above, we modify the objective function in NCA. Besides the neighbor assignment consideration in [27], we also consider the structural consistency between different modalities of features.

$$\begin{gathered}F = \sum_i \sum_{j \in C_i} \frac{exp(-||A_1 x_{ij}||^2)}{\sum\limits_{k \neq i} exp(-||A_1 x_{ik}||^2)} + \sum_i \sum_{j \in C_i} \frac{exp(-||A_2 y_{ij}||^2)}{\sum\limits_{k \neq i} exp(-||A_2 y_{ik}||^2)} + \\ \lambda (\sum_{i,j} exp(-||A_1 x_{ij}||^2)||A_2 y_{ij}||^2 + exp(-||A_2 y_{ij}||^2)||A_1 x_{ij}||^2\end{gathered} \tag{4.19}$$

where x and y are the features in the two different modalities, respectively, A_1 and A_2 are subspaces to be learned for the two views, and λ is a weighting factor between the original NCA objective and the structural consistency. In our problem it is constrained that $\lambda > 0$, which means the structural consistency provide positive contribution in the subspace learning. The different value of λ will be helpful on balancing the two modalities information, and is manually set in our approach.

The objective function in 4.19 is basically composed by three different terms. In the first two terms, the objective is equivalent to performing NCA for two modalities individually. The third term couples the two modalities by exploring the structural consistency. The two factors, $exp(-||A_1 x_{ij}||^2)$ and $exp(-||A_2 y_{ij}||^2)$, are the similarity measurement of i-th and j-th samples in the two views respectively. Considering the item $exp(-||A_1 x_{ij}||^2)||A_2 y_{ij}||^2$, the distance computed in subspace A_2 of the y feature space is weighted by their corresponding subspace distance in x feature space. Therefore, the additional term means that the distance of two similar samples in one modality should be small in the other. It can also be explained with a graph perspective, i.e., the labels of samples in one view should be smooth on the graph in the other view. The two modality cannot be learned independently so a co-learning is required to optimize the objective function.

The objective function, F, is differentiable to both A_1 and A_2 and the results are respectively formulated as follow,

$$\begin{gathered}
\frac{\partial F}{\partial A_1} = 2A_1 \sum_i (p_i \sum_k p_{ik} x_{ik}^T x_{ik} - \sum_{j \in C_i} p_{ij} x_{ij}^T x_{ij}) - \\
\lambda(\sum_{i,j} ||x_{ij}||^2 (exp(-||A_1 x_{ij}||^2)||A_2 y_{ij}||^2 + exp(-||A_2 y_{ij}||^2))) \\
\frac{\partial F}{\partial A_2} = 2A_2 \sum_i (p_i \sum_k p_{ik} y_{ik}^T y_{ik} - \sum_{j \in C_i} p_{ij} y_{ij}^T y_{ij}) - \\
\lambda(\sum_{i,j} ||y_{ij}||^2 (exp(-||A_2 y_{ij}||^2)||A_1 x_{ij}||^2 + exp(-||A_1 x_{ij}||^2)))
\end{gathered} \tag{4.20}$$

We thus adopt a gradient descent approach to solve the optimization problem. One problem in practice is the computation of additional item is quite computational expensive. In order to speed up the gradient computation, the sums that appear in Eq. 4.20 over all samples are truncated. Since the faraway neighbors will not contribute as much as the close neighbors in the objective function, only considering several nearest neighbor of each sample is enough. The Eq. 4.20 can be written as,

$$\begin{gathered}
\frac{\partial F}{\partial A_1} = 2A_1 \sum_i (p_i \sum_k p_{ik} x_{ik}^T x_{ik} - \sum_{j \in N_J^1} p_{ij} x_{ij}^T x_{ij}) - \\
\lambda(\sum_i \sum_{j \in N_J^1} ||x_{ij}||^2 (exp(-||A_1 x_{ij}||^2)||A_2 y_{ij}||^2 + exp(-||A_2 y_{ij}||^2))) \\
\frac{\partial F}{\partial A_2} = 2A_2 \sum_i (p_i \sum_k p_{ik} y_{ik}^T y_{ik} - \sum_{j \in N_J^2} p_{ij} y_{ij}^T y_{ij}) - \\
\lambda(\sum_i \sum_{j \in N_J^2} ||y_{ij}||^2 (exp(-||A_2 y_{ij}||^2)||A_1 x_{ij}||^2 + exp(-||A_1 x_{ij}||^2)))
\end{gathered} \tag{4.21}$$

where N_J^m indicates the first N nearest neighbors in the m-th view. In this thesis we empirically set the size of neighborhood to 20. This is a widely applied approach on reducing computational cost, and the performance will not degrade much as the similarities of samples that have no neighborhood information are usually small. Noted for our problem, it is not applicable to consider every point in the space as a neighborhood. It is prohibitive for such complicated computation.

Besides the improvement of subspace learning, the proposed method have two other advantages. Firstly, only labeled data samples are required in the learning. There is no need to use unlabeled data samples. Secondly, the method is effective even if there are very few labeled samples, which is proved by experiment in the following sections. The main drawback for the proposed solution is the computation complexity.

4.4 Experiments and Performance Evaluation

In order to test the performance of our proposed distance metric co-learning algorithm, we test our methods on audio-visual speaker identification. Besides, to prove the robustness of proposed method, we also implement the algorithm on other applications, including hand-written digit recognition and text analysis. In the following sections we will present the simulation result respectively. Given the feature set A and B, we are doing multi-modality co-learning in our experiment. For result comparison, we compare our approach with the following methods:

(1) k-NN with only feature set A using Euclidean distance;

(2) k-NN with only feature set B using Euclidean distance;

(3) k-NN with feature set A and B using Euclidean distance, i.e., fusing the Euclidean distances in different set of feature space and then performing k-NN;

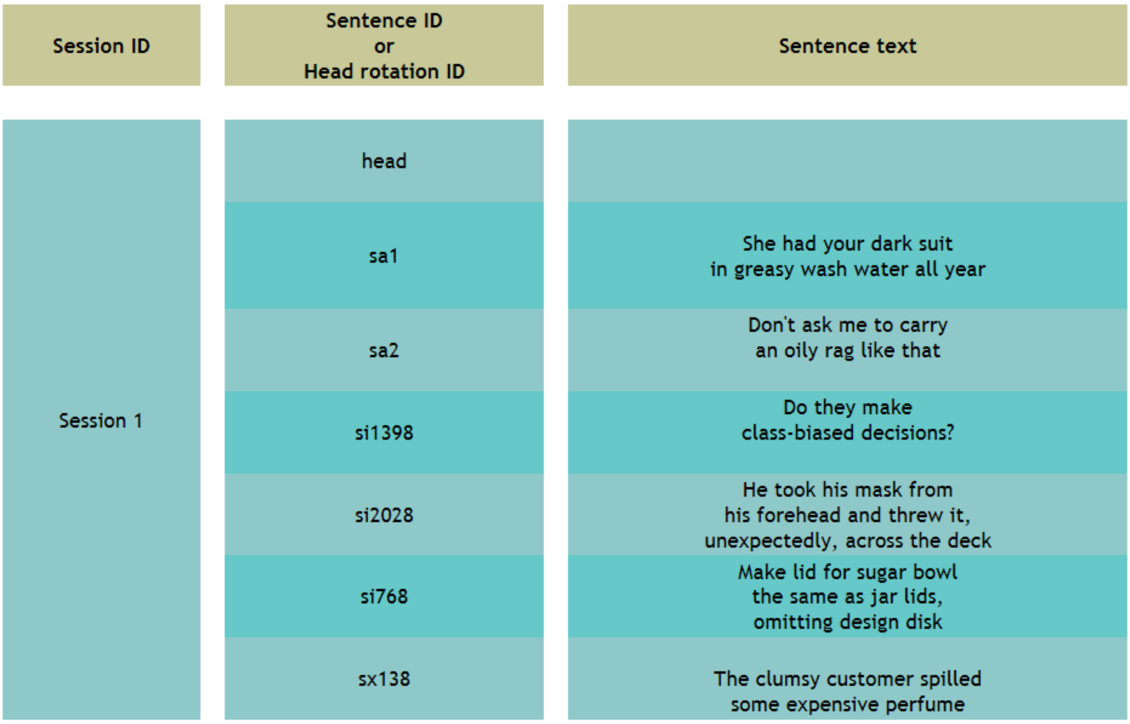
(4) k-NN with two group of features using the distance metrics learned via NCA, i.e., replacing the Euclidean distance with the metrics learned via NCA for each view in the 3rd method.

(5) k-NN after performing CCA on the two set of features.

For simplicity, these five methods are denoted by “Feature A”, “Feature B”, “Fusion+Euclidean”, “Fusion+NCA” and “Fusion+CCA”, respectively.

4.4.1 Audio-Visual Speaker Identification

In this section we apply our method for audio-visual speaker identification on VidTIMIT [70] dataset, which has been widely used, such as in [17] [2]. The VidTIMIT dataset contains 43 persons reciting different sentences, samples are provided in Fig. 4.7. The appearance for each speaker during the speaking is recorded as pictures, as shown in Fig. 4.8. The task is to identify speaker in the query clip with a labeled training set.



Session ID	Sentence ID or Head rotation ID	Sentence text
Session 1	head	
	sa1	She had your dark suit in greasy wash water all year
	sa2	Don't ask me to carry an oily rag like that
	si1398	Do they make class-biased decisions?
	si2028	He took his mask from his forehead and threw it, unexpectedly, across the deck
	si768	Make lid for sugar bowl the same as jar lids, omitting design disk
	sx138	The clumsy customer spilled some expensive perfume

Figure 4.7. Sample of sentences for speakers

In the VidTIMIT Audio-Visual speaker dataset, the audio signal and visual images are treated as the two sets of features for multi-modality learning. According to [17], standard 36-dimensional MFCC [21] are extracted every 10ms over a 20ms sliding window, and they concatenated over a window of 440ms centered on the current frame. Totally 1,584 dimensional audio features are obtained in this way. The data size is too large to compute distance. Therefore a dimensionality reduction is applied by PCA and finally 100 dimensions are preserved for future processing.

For visual images, features are extracted directly by using the pixel values in each frame row by row. Each frame is represented as a vector. A PCA is applied to reduce the number of dimensions in order to speed up the computation. We maintain the 100 dimensions for visual features as well. It is worth noting that our scheme is flexible and we can

Figure 4.8. Example frames in VidTIMIT Dataset

also integrate other audio and visual features which may achieve better performance. The parameter is tuned by 5-fold cross-validation.

In the experiment we randomly select k samples for learning and the others for test, the value of k varies from 3 to 7 due to the dataset limitation. We perform multi-modality distance metric learning using the proposed method and compute the distances between samples. We perform identification by 6 different methods mentioned previously. The result is presented in Fig. 4.9.

As a multimedia pattern recognition problem, audio-visual speaker identification is very challenging. It is easily found in Fig. 4.9 that the Euclidean distance failed to have good performance in such challenging work, some special subspace learning technique is necessary to achieve better measurement of similarity. Multi-modality solutions, both our proposed and CCA, outperformed the one modality version NCA. Our solution is also better than CCA for about 1 or 2 percentage in recognition accuracy.

4.4.2 Other Applications

To prove the robustness of the proposed method, we also implement our proposed method on other applications, such as handwritten digit recognition and text analysis. We present the related experiment result in this section.

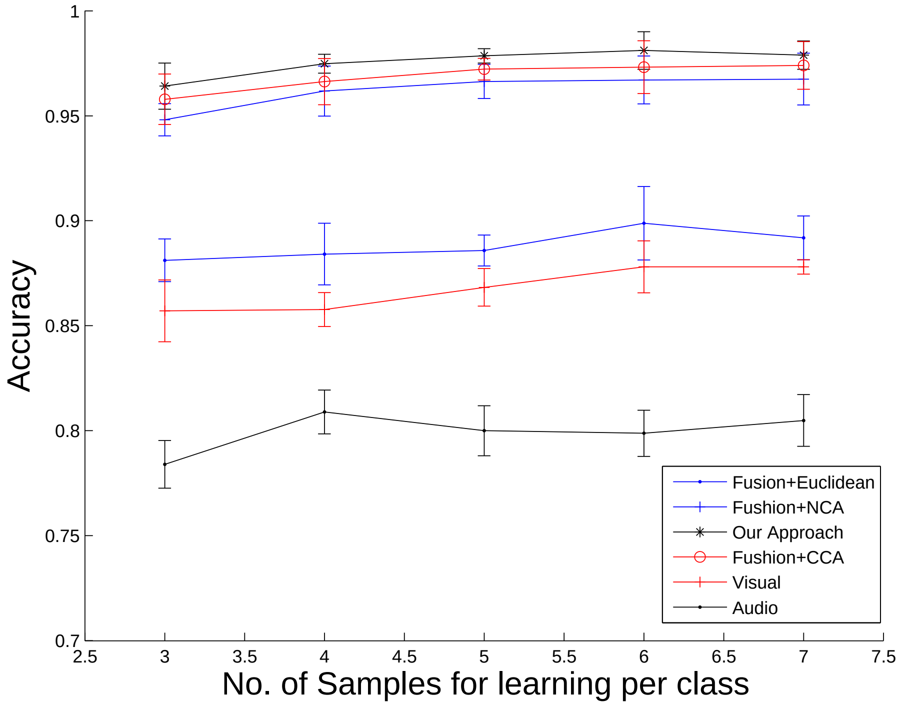


Figure 4.9. Performance on VidTIMIT Audio-Visual dataset

4.4.2.1 Handwritten Digit Recognition

In this section we implement our method on a multi-modality handwritten digit recognition dataset [5]. The handwritten digit recognition is chosen for experiment because it is a relatively simple computer vision work. The input image consists of black or white pixels, the digits are therefore well-separated from the background. There are only ten output categories, which is acceptable in the classification literature. The problem deals with objects in a real two-dimensional space and the mapping from image space to category space has both considerable regularity and complexity.

The dataset is constructed from the UCI machine learning Repository [85]. It is composed by 44 writers and 250 samples from each of them and all the numbers are divided into 10 classes, from 0 to 9 Arabic numerals. All of the image pictures are in binary format, i.e., 0 or 1 for white and black. Some samples of digit are shown in Fig. 4.10.

Two different set of features are extracted respectively from the original digit figure:

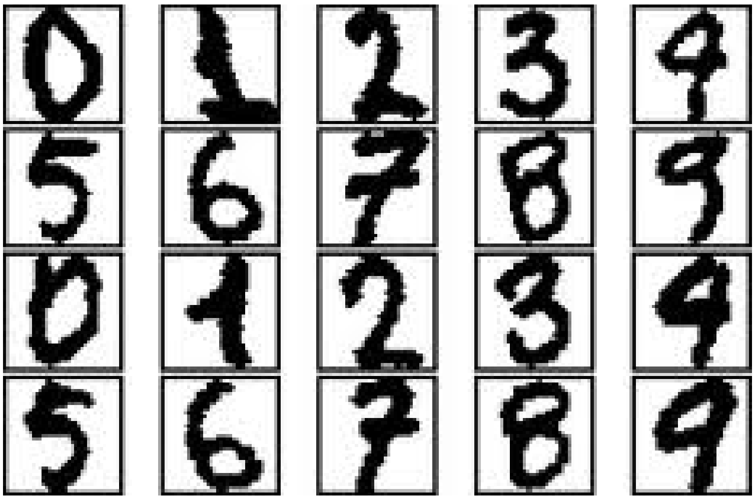


Figure 4.10. Digit Samples in UCI Dataset

a set of pen features and a set of optical features. The 500x500 pixel resolution digit is first normalized to avoid the factor of scaling and rotation. 16-dimensional pen features are extracted to represent the coordinate information at special time interval. This group of feature is named as “pen-based features”. Then the digit is divided into an 8x8 non-overlapping subblocks, and a 64-dimensional optical feature vector is generated by accumulating the pixel values in each block, which is called “optical-based features”.

In the experiment we randomly select k samples from each class for training and use the other samples for test. The value of k varies from 2 to 12 due to the size of dataset. We test 6 different methods on the dataset and present the result in Fig. 4.11. We run 10 times for each experiment and provide the average recognition accuracy and the variance in the figure. From the results we can see that recognition result obtained from single modality, either pen-based or optical-based, is not achieving as good performance as multi-modality solutions. Jointly considering the modalities will certainly enhance the recognition accuracy by considering the structural consistency. Besides, Euclidean distance is not a good evaluation for similarity between samples compared with subspace distance. As a solution considering both multi-modalities and subspace, CCA offers a better recognition result. The distance in CCA subspace is more reliable. Our proposed approach consistently outperforms

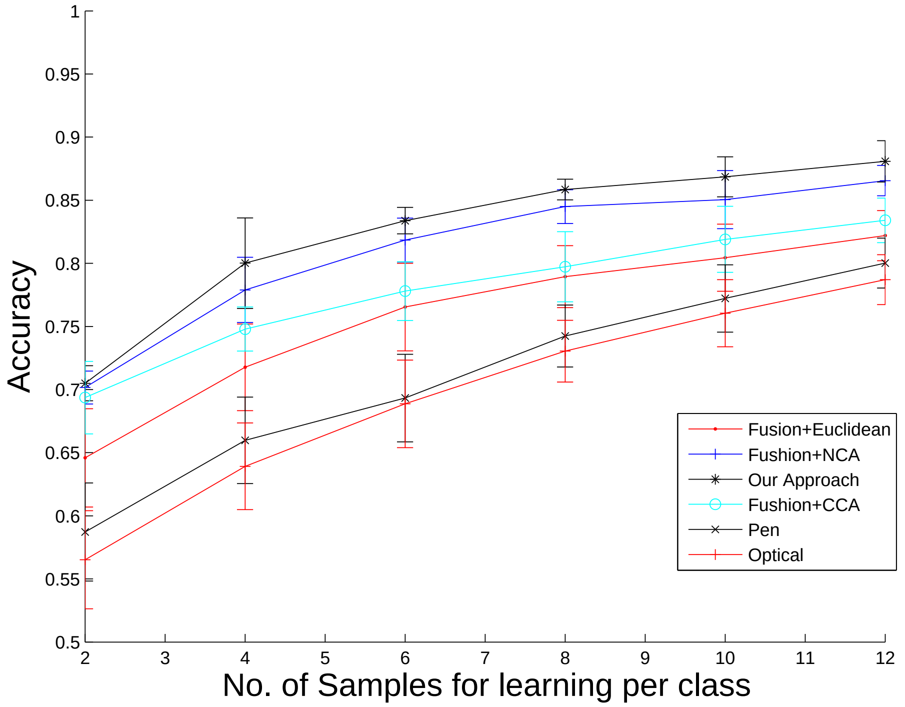


Figure 4.11. Performance on Handwritten Digit Recognition

the other methods. The variance of the proposed method is comparable with other solutions, which means the recognition performance is stable. This confirms the effectiveness of our multi-modality distance metric learning method.

4.4.2.2 Text Content Analysis

In this section we apply our proposed method on the 3 source dataset [81], for text content analysis and classification. Text content analysis, sometimes alternatively referred to as text mining, refers to the process of deriving high-quality information from large paragraph of text. The information is typically derived from statistical pattern learning or within the structured data.

This set is composed by stories from different media. The stories can be categorized as one of the six classes, *Business*, *Entertainment*, *Health*, *Politics*, *Sports* and *Technology*. The text content analysis target on efficiently classify a query text into one of the six the pre-defined categories, with feature extracted from the content.

In the experiment we use the frequency of words as feature. The number of appearance for each word is accumulated into a histogram and a vector is formed to represent the whole story. Specially, to reduce the computation, those words which have no more than 3 appearances in the text are removed from the histogram. Two set of features can be obtained from two different media, and in the experiment we select 250 stories, reported by BBC and Guardian respectively.

Intuitively, the story with same class label should have similar word distribution, especially should have similar key words appearance. So the similarity between two text paragraphs can be found by matching their word histogram. Query story are classified into the one with largest similarity.

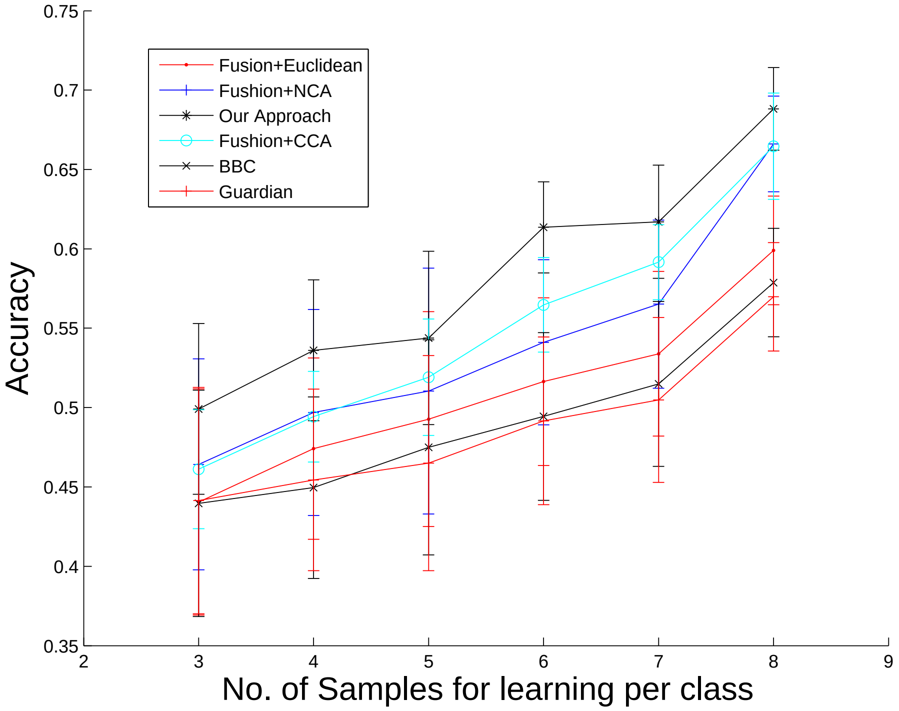


Figure 4.12. Performance on 3 sources text dataset

Similarly as digit recognition, in the experiment we randomly select k stories from each classes for training and distance metric co-learning, and the others are used for test.

The parameter k varies from 2 to 8 due to the limitation of dataset size. We apply 6 different methods mentioned previously and plot the result in Fig. 4.12. From the result we can see the multi-modality-based methods is performing better than single-modality solutions. Also, Euclidean distance is not providing as good measurement for similarity as subspace distance. CCA and our proposed approach are generally performing better than only using the single set of features or Euclidean distance. Our proposed method also outperformed NCA and CCA for about 5 percent and 3 percent, respectively.

4.5 Summary

In this chapter, we propose an multi-modality distance metric co-learning method. The proposed method can be applied for any task with two set of features, i.e., two modalities. The proposed method only requires labeled data, and work even if there is only a few samples. In this chapter we apply the method for digit recognition, text content analysis and audio-visual speaker identification. We learn distance metrics for both set of features. The metric learning scheme not only constructs distance measures based on the label information of training data but also the structural consistency of different modalities. In this way, more reliable metrics can be learned in comparison with learning the two metrics separately. In the experiment we find that multi-modality solutions always performs better than single modality, and solution with subspace distance performs better than Euclidean distance. By considering the structural consistency between modalities, our proposed method outperforms the NCA and CCA in all the three applications. Extensive experiments have demonstrated the effectiveness of our approach.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Video content analysis and understanding are highly demanded for various online application nowadays. There are already lots of related research work and prototype in the literature. The main problem for video processing is that the large amount of data have a lot of redundancy, and the statistical information for understanding is difficult to be detected. On the other hand, for content analysis problem, there are only several potential answers, which can be easily described. The main challenge in such problem is how to effectively and extract the useful information from large amount of data, especially in a very short time for online applications.

In this thesis, we investigated the video-based pattern recognition in detail and improve the recognition accuracy by proposing video representation methods, spatio-temporal modeling solutions and classification approaches. In feature extraction and video representation, we target on finding a comprehensive representation for each video and achieve the reduction in data amount together with preserving the critical information. For spatio-temporal modeling, we proposed differential computation, global subspace learning and local appearance space indexing. For classification, we propose a multi-modality distance metric co-learning approach, which utilize the structural consistency between multiple set of features. Each of these proposed approaches improved the performance either by enhance the recognition accuracy or reducing the computation cost.

- For the global feature extraction and video representation, we detect the visual information from each pixel value. However, millions of pixels exists in one video frame and it is infeasible to process all of them. Down sample is applied to reduce the number of pixels and remove the image noises, followed vectorization of video frame and projection to a high dimensional space. In this way each video clip is represented as a trajectory in the space. We then use PCA to globally reduce the number of dimensions for further simplify the data to be processed together with preserving the most important information for each frame. Similar statistical data are found in similar video sequences, which is embedded in the trajectory. To solve the different duration problem, we try to align the trajectories of different length with two approaches, DLFT and LAPD separately. DLFT focused on the temporal information and use the differential frame as the main discriminative information for classification. LAPD utilized the spatial appearance information to effectively matching the different trajectories by finding the optimal offset for distance computation. Simulation result proved the effectiveness of both approaches, together with the robustness by using multiple dataset.

- For the spatio-temporal modeling part, we extend our global approaches into more detail. It is observed from some special cases that local information should not by skipped in the processing, due to the large contribution for recognition accuracy. However, the local feature computation is basically time consuming. This results in a trade-off between the speed and performance. We proposed two schemes for local indexing, LLAPD and LWAPD. Localization is achieved by partitioning the whole feature space into small sub-blocks and a kd-tree is constructed to describe the local information. Different from the global method, in LLAPD each of the sub-block has a unique subspace to compute the similarity between trajectories. For temporal consideration, a more flexible LWAPD is proposed based on LAPD, with potentially warping operations. The limitation of one-to-one point matching in LAPD is removed and better relationship can be found between trajectories.

- For the effective classification and machine learning approaches, we target on improving the recognition accuracy by utilizing multiple set of features, such as audio and visual features respectively. Traditionally, features from different set are projected into different feature space, and the neighbor information is utilized to describe a similarity within the feature space. With multiple set of features, we construct a correlation between different feature space by using mutual neighbor information. A sample can have corresponding point in different feature space and the locations in one space are helpful to correct the inaccurate description in the other. Distance is computed as an evaluation of similarity between samples in our experiment. It is proved that distance metric learned from multiple feature set is more reliable than from single feature set. And the numerical result is also competitive with some popular approaches in the literature, such as CCA.

The proposed schemes are tested by different applications including video-based human action recognition, gesture recognition and audio-visual speaker identification. However it is not limited in such applications since the algorithm proposed can be generally applied for other video-based pattern recognition tasks. The exploration of our proposed methodology can create a comprehensive understanding that improves upon the state-of-the-art. Numerical result is competitive or better than the ones in the literature. Robustness of the proposed schemes is also demonstrated by result from various dataset with different challenges.

5.2 Future Work

The work presented in this thesis can be extended in different aspects in the future. We summarize some potential directions as follow.

- First, for our proposed global feature extraction and video representation, there is still room for improvement in subspace learning. Various kinds of classifiers could be used to replace the current GMM and KNN one, such as SVM, in DLFT solution. The timing issue is mainly determined by the length of transformation, so it is an interesting problem to exploit the trade-off between the processing time and accuracy. For LAPD approach, there are already two extended versions, but it still has other operations, such as combining the modeling problem with some classical modeling, such as HMM.

- Second, for LLAPD and LWAPD approaches, there is also a trade-off between the performance and time complexity. In LLAPD, more levels in the tree structure leads to slower processing, while in LWAPD, more flexibility cost more time.
- Third, at the multi-modality distance metric co-learning topic, there can be more efficient solutions by optimizing the parameters. Currently the speed is still acceptable, but it may become a drawback when handling larger scale video data. Besides, we only implement two-modality case, which can be extended to three modalities or even more.
- Finally, a possible research direction apply our proposed framework in other problems, such as video searching and retrieval.

REFERENCES

[1] J. K. Aggarwal and Q. Cai, Human Motion Analysis: A Review, *Computer and Image Understanding*, Vol. 73, No. 3, pp.428-440, 1999.

[2] P.S. Aleksic, A.K. Katsaggelos. Audio-Visual Biometrics, *Proceedings of the IEEE*, Vol. 94, No. 11, pp.2025-2044,2006.

[3] C. Bahlmann, B. Hasdonk and H. Burkhardt, On-Line handwriting recognition with support vector machine: a kernel approach. In *Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*, pp. 490-495, 2002.

[4] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, Vol. 19, No. 7, pp. 711-720, 1997.

[5] M. Bilenko and S. Basu and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of 21st International Conference on Machine Learning (ICML)*. pp.81-88, 2004.

[6] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, Actions as space-time shapes. In *Proceedings of 10th IEEE International Conference on Computer Vision (ICCV)*, pp. 1395-1402, 2005.

[7] A. Blum and T. Mitchell, Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 92-100, 1998.

[8] A. F. Bobick and J. Davis, An Appearance-Based Representation of Action, *Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp.307-312, 1996.

[9] A. F. Bobick and J. Davis, The Recognition of Human Movement using Temporal Templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 23, No. 3, pp. 257-267, 2001.

[10] U. Brefeld and T. Scheffer. Co-em support vector learning. In *Proceedings of International Conference on Machine Learning(ICML)*, 2004.

[11] VCA usage increase in British Security, BSIA report.

[12] C. Burges, A Tutorial on Support Vector Machines for Pattern Recogntion, *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121-167, 1998.

[13] L. Campbell, D. Becker, A. Azarbayejani, A. Bobick and A. Pentland, Invariant features for 3d gesture recognition, *Proceedings of international workshop on automatic face and gesture recognition,* pp.157-162, 1996.

[14] J.P. Campbell, Speaker Recognition: A Tutorial. *Proceedings of the IEEE,* Vol. 85, Issue 9, pp. 1437-1462, 1997.

[15] P. Canotilho and R. P. Moreno, Detecting luggage related behaviors using a new temporal boost algorithm. In *Proceedings of IEEE International Workshop on Performance Evaluation of Tracking Surveillance*, pp. 1-6, 2007.

[16] D. B. D. Cao, O. Masoud, and N. Papanikolopoulos, Online motion classification using support vector machines. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2291-2296, 2004.

[17] K. Chaudhuri, S. M. Kakade, K. Livescu and K. Sridharan. Multi-View Clustering via Canonical Correlation Analysis. In *Proceedings of International Conference on Machine learning (ICML)*, pp. 129-136, 2009.

[18] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, A Fully Automated Content-Based Video Search Engine Supporting Spatio-Temporal Queries. *IEEE Transaction on Circuits and Systems for Video Technology (CSVT)* , vol. 8, no. 5, pp. 602-615, September 1998.

[19] S. Cherla, K. Kulkami, A. Kale, V. Ramasubramanian, Towards Fast, View-Invariant Human Action Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Wordshops (CVPRW)*, 2008.

[20] C. Cedras and M. Shah, Motion based Recognition: A Survey. *Image and Vision Computing*, Vol. 13, No. 2, pp. 129-155, 1995.

[21] S. B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28, No. 4, 1980.

[22] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin, 1997.

[23] A. Efros, A. Berg, G. Mori and J. Malik, Recognition Action at a Distance, In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 726-733, 2003.

[24] M. Fleischman, P. Decamp, and D. Roy, Mining temporal patterns of movement for video content classification. In *Proceedings of 8th ACM International Workshop Multimedia Information Retrieval (MIR)*, pp. 183-192, 2006.

[25] N. Ghanem, D. DeMenthon, D. Doermann, and L. Davis, Representation and recognition of events in surveillance video using Petri nets. In *Proceedings of International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pp. 112-121, 2004.

[26] C. Godin and P. Lockwood, DTW schemes for continuous speech recognition, a unified view, *Computer Speech and Language*, Vol. 3, No. 2, pp. 169-198, 1989.

[27] J. Goldberger; S. Roweis; G. Hinton; and R. Salakhutdinov, Neighbourhood components analysis. In *Advances in Neural Information Processing System (NIPS)*. pp. 513-520, 2005.

[28] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as Space-Time Shapes , *IEEE Transaction on Pattern Analysis and Machine Intelligence(PAMI)*, vol. 29, no.12, pp. 2247-2253, 2007.

[29] D. Hardoon, S. Szedmak and J. Shawe-Tayler, Canonical Correlation Analysis: An overview with application to Learning Methods, *Neural Computation*, Vol.16, No. 12, pp. 2639-2664, 2004.

[30] T. Hastie and W. Stuetzle, Principal curves, *Journal of the American Statistical Association*, Vol.84, pp. 502-516, 1989.

[31] X. He, and P. Niyogi, Locality Preserving Projections, In *Advances in Neural Information Processing Systems(NIPS)*, 2003.

[32] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, RASTA-PLP speech analysis technique. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp.121-124, 1992.

[33] J. Hoey and J. Little, Representation and Recognition of Complex Human Motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1752-1759, 2000.

[34] S.C.H. Hoi, L. Wei, M.R. Lyu, W. Ma; Learning Distance Metrics with Contextual Constraints for Image Retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2072-2078, 2006.

[35] S. Hongeng and R. Nevatia, Multi-agent event recognition. In *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 84-93, 2001.

[36] H. Hotelling, Relations between two sets of variants. *Biometrika*, Vol.28, pp. 321-377, 1936.

[37] R. Hu, R. I. Damper, Fusion of two classifiers for speaker identification: removing and not removing silence. In *Proceedings of International Conference on Information Fusion*, pp. 429-436, 2005.

[38] J. Hu, B. Ray and L. Han, An Interweaved HMM/DTW Approach to Robust Time Series Clustering. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, pp. 145-148, 2006.

[39] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 166-173, 2005.

[40] B. Kegl, A. Krzyzak, T. Linder, and K. Zeger, Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol.22, no.3, pp.281-297, 2000.

[41] B. Kegl and A. Krzyzak, Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, vol. 24, no. 1, pp. 59-74, 2002.

[42] T. Kim, S. F. Wong and R. Cipolla, Tensor Canonical Correlation Analysis for Action Classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[43] T-K. Kim, J. Kittler, and R. Cipolla. Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, Vol.29, No.6, pp. 1005-1018, 2007.

[44] I. Laptev and T. Kindeberg, Space-time interest points. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 432-439, 2003.

[45] I. Laptev and P. Perez, Retrieving actions in movies. In *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 1-8, 2007.

[46] I. Laptev, M. Marszalet, C. Schmid and B. Rozenfeld, Learning Human Actions from Movies. In *IEEE Conference on Computer Vision and Patter Recotnition (CVPR)*, 2008.

[47] G. Lavee, E. Rivlin, M. Rudzsky, Understanding video events: a survey of methods for automatic interpretation of semantic occurences in video. *IEEE Transactions on systems, man, and cybernetics, Part C: applications and reviews (TSMC-C)*, vol. 39, No. 5, pp.489-504, 2009.

[48] B. Legrand, C. Chang, S. Ong, S. Neo and N. Palanisamy, Chromosome classification using dynamic time warping. *Pattern Recognition Letters*, Vol.29, Issue 3, pp. 215-222, 2008.

[49] Z. Li, Y. Fu, S. Yan, and T.S. Huang, Real-Time Human Action Recognition by Luminance Field Trajectory Analysis. In *ACM Multimedia*, pp.671-675, 2008.

[50] J. Liu, J. Luo and M. Shah, Recognizing Realistic Actions from Videos ”in the Wild”. In *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 1996-2003, 2009.

[51] J. Liu, Y. Yang and M. Shah, Learning Semantic Visual Vocabularies using Diffusion Distance. In *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 461-468, 2009.

[52] K. Livescu, M. Stoehr, Multi-view Learning of Acoustic Features for Speaker Recognition. In *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009.

[53] D. G. Lowe, Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, Vol. 60, No. 2, pp. 91-110, 2004.

[54] A. Marzal and V. Palaz, Dynamic time warping of cyclic strings for shape matching, *Pattern Recognition and Image Analysis, Springer*, pp.644-652, 2005.

[55] D. Minnen, T. Westeyn, and T. Starner, Recognizing soldier activities in the field, In *International Workshop Wearable Implantable Body Sensor Networks*, pp. 236-241, 2007.

[56] D. Minnen, T. Starner, I. Essa, and C. Isbell,Improving activity discovery with automatic neighborhood estimation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2814-2819, 2007.

[57] R. Muscillo, S. Conforto, M. Schmid, P. Caselli and T. D. Alessio, Classification of motor activities through derivative dynamic time warping applied on accelerometer data. In *Proceedings of 29th IEEE International Conference on Engineering in Medicine and Biology Society*, pp. 4930-4933, 2007.

[58] J.C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words, *British Machine Vision Conference (BMVC)*, pp. 1249-1258, 2006.

[59] K. Nigam, and R. Ghani, Analyzing the Effectiveness and Applicability of Co-training. In *Proceedings of 9th International Conference on Information and Knowledge (CIKM)*, pp.86-93, 2000.

[60] S. Nowozin, G. Bakir, and K. Tsuda, Discriminative subsequence mining for action classification. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 1919-1923, 2007.

[61] T. Oates, M. Schmill and P Cohen, A method for clustering the experiences of a mobile robot that accords with human judgments. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pp 846-851, 2000.

[62] A. Oikonomopoulousm, I. Patras and M. Pantric, Spatiotemporal Saliency for Human Action Recognition. In *IEEE International Conference on Multimedia and Expo.(ICME)*, pp. 430-433, 2005.

[63] N. Oliver, B. Rosario, and A. Pentland, A Bayesian computer vision system for modeling human interactions. *IEEE Transaction on Pattern Analysis and Machine Intellegence (PAMI)*, Vol. 22, No. 8, pp. 831-843, 2000.

[64] C. Piciarelli, G. Foresti, and L. Snidaro, Trajectory clustering and its applications for video surveillance. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 40-45, 2005.

[65] M. Pierobon, M. Marcon, A. Sarti, S. Tubaro, Clustering of Human Actions Using Invariant Body Shape Descriptor and Dynamic Time Warping. In *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 22-27, 2005.

[66] M. Pittore, C. Basso, and A. Verri, Representing and recognizing visual dynamic events with support vector machines. In *Proceedings of International Conference on Image Analaysis and Processing (ICIAP)*, pp. 18-23, 1999.

[67] P. Ribeiro, P. Moreno, and J. S. Victor, Boosting with temporal consistent learners: An application to human activity recognition. In*Proceedings of International Symposium on Visual Computing*, pp. 464-475, 2007.

[68] M. D. Rodriguez, J. Ahmed, and M. Shah, Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[69] H. Sakoe and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transanctions on Acoustics, Speech, and Signal Process*, Vol. 26, pp.43-49, 1978.

[70] C. Sanderson. Biometric Person Recognition: Face, Speech and Fusion. *VDM-Verlag*, 2008.

[71] B. Sarker and K. Uehara, Efficient parallelism for mining sequential rules in time sereis data: a lattice based approach. *International Journal of Computer Science and Network Security*, Vol. 6, No. 7A, pp. 137-143, 2006.

[72] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *IEEE International Conference on Pattern Recognition (ICPR)*,2004.

[73] E. Shechtman and M. Irani, Space-time behavior based correlation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[74] V. Sindhwani, P. Niyogi, and M. Belkin, A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 824-831, 2005.

[75] V. Sindhwani and D. S. Rosenberg, An RKHS for multi-view learning and manifold co-regularization. In *Proceedings of International Conference on Machine Learning (ICM-L)*, pp. 976-983, 2008.

[76] S. M. Smith and J. M. Brady. ASSET-2: Real-time motion segmentation and shape tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol.17, No.8: pp. 814-820, 1995.

[77] P. Smith, N. da Vitoria Lobo, and M. Shah, Temporal boost for event recognition. In *Proceedings of IEEE International Conference on Computer Vision(ICCV)*, pp. 733-740, 2005.

[78] C. G. M. Snoek, M. Worring, A. Smeulders, Early versus late fusion in semantic video analysis. In *ACM Multimedia*, pp. 399-402, 2005.

[79] A. Sundaresan, A. R. Chowdhury, R. Chellappa. A hidden Markov model based framework for recognition of humans from gait sequences. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Vol. 2, pp. 93-96, 2003.

[80] D. Tao, X. Li, X. Wu, S. J. Maybank, General Tensor Discriminant Analysis and Gabor Features for Gait Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 10, 1700-1715, 2007.

[81] [online]http://mlg.ucd.ie/datasets/3sources.html.

[82] S. Tran and L. S. Davis, Event modeling and recognition using Markov logic networks. In *Proceedings of Europe Conference on Computer Vision (ECCV)*, pp. 610-623, 2008.

[83] P. Turaga, R. Chellappa, V. S. Subrahmanian and O. Udrea, Machine Recognition of Human Activities: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, Vol. 18, No. 11, pp. 1473-1488, 2008.

[84] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, vol. 3, no.1, 1991.

[85] [online]http://archive.ics.uci.edu/mml/

[86] D. Vail, M. Veloso, and J. Lafferty; Feature selection in conditional random fields for activity recognition. In *IEEE International Conference on Intelligent Robots and Systems*, pp.3379-3384, 2007.

[87] K.Q. Weinberger, J. Blitzer, L.K. Saul, Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing System(NIPS)*, 2005.

[88] D. Weinland, E. Boyer, R. Ronfard, Action Recognition from Arbitrary Views using 3D Exemplars, *International Conference on Computer Vision (ICCV)*, pp. 1-7, 2007.

[89] Y. Wu, E. Y. Chang, K. C. Chang and J. Smith, Optimal multimodal fusion for multimedia data analysis, In *ACM Multimedia*, pp. 572-579, 2004.

[90] Y. Wu and T. Yu, A Field Model for Human Detection and Tracking. *IEEE Transections on Pattern Analysis and Machine Intelligence (PAMI)*, Vol.28, No.5., pp.753-765, 2006.

[91] Y. Xie, B. Wiltgen, Adaptive Feature Based Dynamic Time Warping, *International Journal of Computer Science and Network Security*, vol. 10, No. 1, pp. 264-273, 2010.

[92] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell, Distance Metric Learning with Application to Clustering with Side-information. *Advances in Neural Information Processing Systems (NIPS)*, pp.505-512, 2002.

[93] D. Xu and S. F. Chang, Visual event recognition in news video using kernel methods with multi-level temporal alignment. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2007.

[94] O. Yakhnenko, V. Honavar; Multiple label prediction for image annotation with multiple Kernel correlation models. In *IEEE on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 8-15, 2009.

[95] J. Yamato, J. Ohya, and K. Ishii, Recognizing Human Action in Time Sequential Images Using Hidden Markov Model. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1992.

[96] P. Yan, S. M. Khan, and M. Shah, Learning 4D Action Feature Models for Arbitrary View Action Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[97] M. H. Yang, Face Recognition Using Kernel Methods, *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[98] Y. Yang, J. Liu, and M. Shah, Video Scene Understanding Using Multi-scale Analysis, *Proceedings of International Conference on Computer Vision (ICCV)*, pp.1669-1676, 2009.

[99] J. Ye; Z. Zhao; H. Liu; Adaptive Distance Metric Learning for Clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-7, 2007.

[100] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, vol. 1, pp.150-157, 2005.

[101] Y. Yuan, H. Zheng, Z. Li, D. Zhang, Video action recognition with spatio-temporal graph embedding and spline modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2422-2425, 2010.

[102] M. Faundez-Zanuy, On-line signature recognition based on VQ-DTW, *Pattern Recognition*, Vol. 40, Issue 3, pp. 981-992, 2006.

[103] L. Zelnik-Manor and M. Irani, Event-based analysis of video. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 123-130, 2001.

[104] L. Zelnik-Manor and M. Irani, Statistical analysis of dynamic actions, *IEEE Transactions on Pattern Analysis and Machine Intellegence (PAMI)*, Vol. 28, No. 9, pp. 1530-1535, Sep. 2006.

[105] H. Zheng, Z. Li, Y. Fu, Human Action Recognition with Luminance Field Trajectory Projection and Alignment. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 842-845, 2009.

[106] H. Zheng, Z. Li, A. K. Katsaggelos and J. You, Indexed Spatio-Temporal Appearance Models for Query-driven Video Action Recognition. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2011.

[107] H. Zheng, M. Wang, Z. Li, Audio-visual Speaker Identification with Multi-view Distance Metric Learning. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 4561-4564, 2010.

[108] H. Zheng, Z. Li, Y. Fu, A. K. Katsaggelos and J. You, Video Activity Recognition by Luminance Differential Trajectory and Aligned Projection Distance, accepted by *Handbook on Statistics*, 2012.

[109] H. Zheng, Z. Li, Y. Yuan, A. K. Katsaggelos and J. You, Video-based Human Action Recognition by Dynamic Wrapped Local Spatio-temporal Indexing, to be submitted to *IEEE Transactions on Information and Forensics and Security (TIFS)*, 2012.