ABSTRACT

Title of dissertation: Learning Visual Classifiers
From Limited Labeled Images

Jaishanker K. Pillai, Doctor of Philosophy, 2013

Dissertation directed by: Professor Rama Chellappa
Department of Electrical and Computer Engineering

Recognizing humans and their activities from images and video is one of the key goals of computer vision. While supervised learning algorithms like Support Vector Machines and Boosting have offered robust solutions, they require large amount of labeled data for good performance. It is often difficult to acquire large labeled datasets due to the significant human effort involved in data annotation. However, it is considerably easier to collect unlabeled data due to the availability of inexpensive cameras and large public databases like Flickr and YouTube. In this dissertation, we develop efficient machine learning techniques for visual classification from small amount of labeled training data by utilizing the structure in the testing data, labeled data in a different domain and unlabeled data.

This dissertation has three main parts. In the first part of the dissertation, we consider how multiple noisy samples available during testing can be utilized to perform accurate visual classification. Such multiple samples are easily available in video-based recognition problem, which is commonly encountered in visual surveillance. Specifically, we study the problem of unconstrained human recognition from iris images. We develop a Sparse Representation-based selection and recognition scheme, which learns the underlying structure of clean images. This learned structure is utilized to develop a quality measure, and a quality-based fusion scheme is proposed to combine the varying evidence. Furthermore, we extend the method to incorporate privacy, an important requirement in

practical biometric applications, without significantly affecting the recognition performance.

In the second part, we analyze the problem of utilizing labeled data in a different domain to aid visual classification. We consider the problem of shifts in acquisition conditions during training and testing, which is very common in iris biometrics. In particular, we study the sensor mismatch problem, where the training samples are acquired using a sensor much older than the one used for testing. We provide one of the first solutions to this problem, a kernel learning framework to adapt iris data collected from one sensor to another. Extensive evaluations on iris data from multiple sensors demonstrate that the proposed method leads to considerable improvement in cross sensor recognition accuracy. Furthermore, since the proposed technique requires minimal changes to the iris recognition pipeline, it can easily be incorporated into existing iris recognition systems.

In the last part of the dissertation, we analyze how unlabeled data available during training can assist visual classification applications. Here, we consider still image-based vision applications involving humans, where explicit motion cues are not available. A human pose often conveys not only the configuration of the body parts, but also implicit predictive information about the ensuing motion. We propose a probabilistic framework to infer this dynamic information associated with a human pose, using unlabeled and unsegmented videos available during training. The inference problem is posed as a non-parametric density estimation problem on non-Euclidean manifolds. Since direct modeling is intractable, we develop a data driven approach, estimating the density for the test sample under consideration. Statistical inference on the estimated density provides us with quantities of interest like the most probable future motion of the human and the amount of motion information conveyed by a pose. Our experiments demonstrate that the extracted motion information benefits numerous applications in computer vision. In particular, the predicted future motion is useful for activity recognition, human trajectory synthesis and motion prediction. Furthermore, the estimated amount of motion information in a pose provides a novel criteria for video summarization.

Learning Visual Classifiers From Limited Labeled Images

by

Jaishanker K. Pillai

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Rama Chellappa, Chair/Advisor
Professor Larry Davis
Professor Min Wu
Professor Ramani Duraiswami
Professor Amitabh Varshney

DEDICATION

To my parents and sisters, who have been a great source of inspiration and support through out my life.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Prof. Rama Chellappa for his guidance and mentoring through out my graduate life. I find his lessons extremely valuable not only in professional life, but also in being a better human being. The discussions we had were enlightening and enjoyable, broadening my perspective about research and the field of computer vision. I am also thankful to Prof Chellappa for providing a wonderful research environment, sufficient intellectual freedom to pursue new ideas and constant encouragements to explore difficult problems without worrying about failures.

It is an honor to have Prof. Larry Davis, Prof. Min Wu, Prof. Ramani Duraiswami and Prof. Amitabh Varshney in my dissertation committee. I am thankful to them for serving in my committee and providing insightful suggestions to improve this dissertation. During my student life, I was fortunate to interact with two of the sharpest minds in Machine Learning namely Prof. Haal Daume and Dr. Oncel Tuzel. I cherish the discussions I had with them, which sparked my interest in the field. I would like to thank Dr. Srikumar Ramalingam, who along with Dr. Oncel Tuzel mentored me during my internships in Mitsubishi Electric Research Lab. They played an important part in making me realize the importance of developing fast and efficient algorithms in vision. I am also grateful to Prof Prakash Narayan, Prof Ray Liu, Prof David Jacobs, Prof Uzi Vishkin, Prof K. R. Ramakrishnan and Prof Venu Madhav Govindu for their enlightening lectures during my graduate studies.

I would like to express my gratitude to Dr. Vishal Patel, Prof. Pavan Turaga, Dr. Nalini Ratha, Prof. Ashok Veeraraghavan, Dr. Nitesh Shroff, Prof. Aswin Sankaranarayanan

and Dr. Mahesh Ramachandran for their mentoring through out my Ph.D. It was a pleasure to work with Raviteja Vemulapalli, Maria Puertas, Ashish Srivastava and Jayant Kumar over the last few years. I would also like to thank my labmates and friends including Hien Nguyen, Sima Taheri, Ming Du, Sumit Shekhar, Garrett Warnell, Huy Tho, David Shaw, Senthil Kumar and others for making my graduate life memorable.

I would like to thank the staff in UMIACS, ECE and Cfar. In particular, special thanks to Janice Perone, Dr. Tracy Chung, Melanie Prange and Arlene Schenk.

My parents and family have been a great source of inspiration and support through out my life. This dissertation would not have been possible without the constant support, encouragement and love of my entire family - my parents Dr N. Kamalasanan Pillai and Saraswati, sisters Dr. Sheila Pillai and Dr. Swapna Pillai, brother-in law Sunil and grand father late Prof. Kochukrishna Pillai.

Finally, I would like to thank God Almighty.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Motivation

Supervised Learning techniques have made tremendous contributions to computer vision, leading to the development of robust algorithms. Viola and Jones [4] developed a robust framework for face detection through boosting-based cascade rejection classifier. Pedestrian detection was performed by Dalal *et. al.* [5] by classifying Histogram of Oriented Gradients (HOG) features using Support Vector Machines (SVM). To model articulated human poses, Felzenswab *et. al.* [6] developed discriminative part models based on Latent SVMs. Pose estimation algorithms have been developed by Andriluka *et. al.* [7] using part-based models. Robust algorithms have been proposed for human action recognition using space time interest points and SVMs with Histogram Intersection kernels [8].

While these algorithms have advanced the state-of-the art significantly, their performance is often limited by the amount of labeled training data available. Labeling is expensive and time consuming due to the significant amount of human effort involved. However, collecting unlabeled visual data is becoming considerably easier due to the availability of low cost surveillance cameras and large Internet databases like Flickr and YouTube. This leads us to an interesting question: Can unlabeled or weakly labeled data be used along with small amount of labeled data to develop accurate visual classifiers? We address this question in this dissertation by developing semi-supervised algorithms for visual classifi-

cation tuned to the application in hand.

1.2 Proposed Algorithms and their Contributions

In this section, we briefly describe the algorithms introduced in this dissertation and their key contributions.

1. **Secure and unconstrained iris recognition using Sparse Representations and Random Projections:**

   In the first part of the dissertation, we consider how multiple noisy samples available during testing can be utilized to perform accurate visual classification. Specifically, we study the problem of unconstrained human recognition from iris images. In this problem, while the training images are clean iris templates of subjects, the images during testing often have large amount of acquisition artifacts like motion blur, occlusion, specular reflections and off angle rotation, due to the unconstrained nature of acquisition. However, multiple samples are available as the test subject moves towards the sensor, which is normally part of an access control system. Hence, we propose a Sparse Representation-based selection and recognition scheme, which learns the underlying structure of clean images [9, 10]. The introduced algorithm simultaneously selects the good iris sectors, recognizes them separately and combines the numerous recognition results using a Bayesian Fusion framework. Furthermore, we extend the method to incorporate privacy using Random Projections [11], an important requirement in practical biometric systems, without significantly affecting the recognition performance.

Contributions: The proposed quality measure can handle wide variety of artifacts like occlusion, blur, specular reflections and off angle rotations of the iris image. The introduced quality based fusion scheme is found to produce state-of-the-art results. We also introduce one of the early algorithms for iris recognition from videos. The proposed cancelable scheme incorporates privacy without significantly reducing the recognition accuracy, unlike existing algorithms for the same purpose.

2. **Sensor Adaptation in Iris Recognition:**

   In the second part, we analyze how labeled data in a different domain can aid visual classification. We consider the problem of shifts in acquisition conditions during training and testing, which is very common in biometrics. With the development of new sensors for iris recognition and the improvement of existing ones, enrollment using one sensor and verification with another assumes great relevance. While verifying test samples using data enrolled from a different sensor can often lead to lower accuracy, enrolling subjects every time a new sensor is deployed is expensive and time consuming. We propose one of the first comprehensive solution to this problem, a machine learning technique to efficiently mitigate the cross-sensor performance degradation, by adapting the iris samples from one sensor to another. We developed a novel optimization framework for learning transformations on iris biometrics. We then utilize this framework for sensor adaptation, by reducing the distance between samples of the same class, and increasing it between samples of different classes, irrespective of the sensors acquiring them. Extensive evaluations on iris data from multiple sensors demonstrate that the proposed method leads to considerable improvement in cross sensor recognition accuracy. Furthermore, since

the proposed technique requires minimal changes to the iris recognition pipeline, it can easily be incorporated into existing iris recognition systems.

Contributions: The proposed method is one of the first comprehensive solution for the sensor mismatch problem in iris biometrics. The introduced solution leads to considerable improvement in cross-sensor matching. It is robust to alignment errors, and can also handle real-valued feature representations. The proposed technique is fast, requiring limited changes to the existing iris recognition pipeline. Hence, it can easily be incorporated into existing iris recognition systems. The framework presented in this dissertation, for developing transformations of iris codes having desired properties, can also be utilized for performing numerous tasks in iris biometrics, such as max-margin classification, dimensionality reduction, and metric learning.

3. **Dynamic Inference from Single Images of Humans:**

   In the last part of the dissertation, we analyze how unlabeled data available during training can assist visual classification applications. Here we demonstrate the usefulness of unlabeled videos in still image-based vision applications involving humans. Our work is motivated by the observation that human pose often conveys not only the configuration of the body parts but also possesses predictive information about the ensuing motion. Image-based vision applications which lack explicit motion information can benefit from this implicit information. However, computational algorithms to infer and utilize it in computer vision applications are limited. In this paper, we propose a probabilistic framework to infer the dynamic information associated with a human pose. The inference problem is posed as a non-

parametric density estimation problem on non-Euclidean manifolds. Since direct modeling is intractable, we develop a data driven approach, estimating the density for the test sample under consideration. Statistical inference on the estimated density provides us with quantities of interest like the most probable future motion of the human and the amount of motion information conveyed by a pose. Our experiments demonstrate that the extracted motion information benefits numerous applications in computer vision. In particular, the predicted future motion is useful for activity recognition, human trajectory synthesis and motion prediction. Furthermore, the estimated amount of motion information in a pose provides a novel criteria for video summarization.

Contributions: We explore the potential of the implicit dynamic information conveyed by a human pose. We develop a probabilistic framework to model it. Using this framework, we estimate the amount of dynamic information conveyed by a pose and predict the probable future motion. The proposed method requires limited manual supervision since it uses unlabeled and unsegmented human videos for training, and can easily be implemented. We demonstrate the usefulness of the estimated dynamic information in a variety of vision applications like human motion prediction, activity recognition and video summarization.

1.3 Organization

This dissertation is organized as follows. In Chapter 2, we present the unconstrained iris recognition algorithm using Sparse Representations and Random Projections. Chapter 3

introduces sensor adaptation for iris recognition. Inference of motion information from still images of humans is described in Chapter 4. We conclude the dissertation and discuss future directions in Chapter 5.

Chapter 2

Unconstrained Iris Recognition

Iris recognition is one of the most promising approaches for biometric authentication [12]. Most existing algorithms rely on the fine texture features extracted from the iris for recognition. Hence their performances degrade significantly when the image quality is poor [12, 1]. This seriously limits the application of the iris recognition system in unconstrained scenarios, where the acquired image could be of low quality due to motion, partial co-operation or the distance of the user from the scanner.

In this dissertation, we develop a framework for unconstrained iris recognition. When the acquisition conditions are not constrained, many of the acquired iris images suffer from defocus blur, motion blur, occlusion due to the eyelids, specular reflections and segmentation errors. Fig. 2.1 shows some of these distortions on images from the ICE2005 dataset [1]. However, the images during enrollment are clean images with limited artifacts, since they are acquired under controlled settings. Hence we need to develop a algorithms for iris recognition using the small amount of labeled samples with limited distortions, to handle test samples with significant acquisition artifacts. However, often multiple images of the subject are available during testing, as the subject moves towards the sensor. This availability of multiple test images can be utilized to improve the recognition performance. In this work, we employ a sparse representation framework [13] to capture the structure of the clean training images and utilize it to estimate the quality of the test

samples. We then utilize a quality-based fusion framework, combine the results of the individual sectors on the iris based on their quality. The proposed method is significantly faster than the original sparse representation approach [13] as it facilitates parallelization and reduces the size of the dictionary size, as will become apparent.

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

Figure 2.1: Some poorly acquired iris images from the ICE dataset [1]. Note that image (a) has specular reflections on the iris and is difficult to be segmented correctly due to the tilt and non circular shape. Images (b) and (d) suffer from blurring, whereas image (c) is occluded by the shadow of the eyelids.

The performance of most existing iris recognition algorithms depends strongly on the effectiveness of the segmentation algorithm. Iris image segmentation normally involves identifying the ellipses corresponding to pupil and iris, and detecting the region inside these ellipses that is not occluded by the eyelids, eyelashes and specular reflections. Unfortunately, in unconstrained scenarios, correctly segmenting the iris images is extremely challenging [14]. The proposed selection algorithm removes input images with poorly segmented iris and pupil ellipses. Furthermore, since the introduced recognition scheme is robust to small levels of occlusions, accurate segmentation of eyelids, eyelashes and specular reflections are no longer critical for achieving good recognition performance.

Another important aspect in iris biometrics is security and privacy of the users. When the texture features of one's iris are stored in a template dictionary, a hacker could possibly

break into the dictionary and steal these patterns. Unlike credit cards, which can be revoked and reissued, biometric patterns of an individual cannot be modified. So, directly using iris features for recognition is extremely vulnerable to attacks. To deal with this, the idea of cancelable iris biometrics has been introduced in [15, 16, 17], which can protect the original iris patterns as well as revoke and reissue new patterns when the old ones are lost or stolen. In this paper, we introduce two methods for incorporating security into the proposed iris recognition system, namely, random projections and random permutations. Our methods can issue a different template for each application based on the original iris patterns of the person, generate a new template if the existing one is stolen while retaining the original recognition performance. The representation prevents extraction of significant information about the original iris patterns from cancelable templates.

Organization of the Chapter: In Section 2.1, we discuss some of the existing algorithms for iris image selection, recognition and cancelability. The theory of sparse representation is summarized in Section 2.2. The Bayesian fusion framework for selecting and recognizing iris images is described in 2.3. We extend our method to video-based iris recognition in section 2.4 and discuss how to handle alignment in Section 2.5. Two schemes for introducing cancelability into our framework are proposed in 2.6. Experiments and results are presented on simulated and real iris images in Section 2.7.

2.1 Related Work

In this section, we briefly describe some of the existing methods for iris recognition, image quality estimation and cancelability.

Iris recognition: The first operational automatic iris recognition system was developed by Daugman [18] in 1993, in which Gabor features were extracted from scale normalized iris regions and quantized to form a 2K bit iris code. The normalized Hamming distance between the iris code of the test and the training iris images was used for recognition. Wildes [19] used Laplacian of a Gaussian filter at multiple scales to produce a template and used the normalized correlation as the similarity measure. In recent years, researchers have analyzed aspects like utilizing real valued features for recognition, developing alternate ways of obtaining the binary codes and combining multiple features. See [12] for an excellent survey of recent efforts on iris recognition.

Several studies have shown that accurate quality estimation can improve the performance either by rejecting the poor quality images or by fusing the quality information during matching [12, 20, 21]. Daugman used the energy of the high frequency components as a measure of blur [18]. Proenca and Alexandre trained a neural network to identify common noise degradations in iris images [22]. Zhu *et al.* used the wavelet coefficients to evaluate the quality of iris images [23]. The Fourier spectra of local iris regions was used by Ma *et al.* to characterize blur and occlusion [24]. With the exception of Daugman's method, these algorithms are specialized for image selection, which requires a separate method for recognizing iris images. Also, these algorithms utilize some property of the iris image to measure image quality and cannot handle the wide variety of common artifacts such as specular reflections and occlusion. In contrast to these methods, the image quality measure introduced in this paper can handle segmentation errors, occlusion, specular reflections, and blurred images. The proposed method also performs both selection and recognition in a single step.

Iris Recognition from Videos : Though research in iris recognition has been extremely active in the past decade, most of the existing results are based on recognition from still iris images [25]. Multiple iris images have been used in the past to improve performance. Du *et al.* [26] demonstrated higher rank one recognition rates by using three gallery images instead of one. Ma *et al.* [27] also enrolled three iris images and averaged the three Hamming distances to obtain the final score. Krischen *et al.* [28] used the minimum of the three Hamming distance as the final score. Schmid *et al.* [29] demonstrated that fusing the scores using log likelihood ratio gave superior performance when compared to average Hamming distance. Liu *et al.* [30], Roy and Bhattacharya [31] used multiple iris images for training classifiers.

The distortions common in iris image acquisition like occlusion due to eyelids, eye lashes, blur, and specular reflections will differ in various frames of the video. So by efficiently combining the different frames in the video, the performance could be improved. Temporal continuity in iris videos was used for improving the performance by Hollingsworth *et al.* [25]. The authors introduced a feature level fusion by averaging the corresponding iris pixels and a score level fusion algorithm combining all the pairwise matching scores. Though averaging reduces the noise and improves the performance, it required images to be well segmented and aligned, which may often not be possible in a practical iris recognition system. We will introduce a quality based matching score that gives higher weight to the evidence from good quality frames, yielding superior performance even when some video frames are poorly acquired.

Cancelable iris biometrics: The concept of cancelable biometrics was first introduced by Ratha *et al.* in [16, 17]. A cancelable biometric scheme intentionally distorts the original

biometric pattern through a revocable and non-invertible transformation. The objectives of a cancelable biometric system are as follows [15]:

- Different templates should be used in different applications to prevent cross matching.
- Template computation must be non-invertible to prevent unauthorized recovery of biometric data.
- Revocation and reissue should be possible in the event of compromise, and
- Recognition performance should not degrade when a cancelable biometric template is used.

In [32], Hash functions were used to minimize the compromise of the private biometric data of the users. Cryptographic techniques were applied in [33] to increase the security of iris systems. In [34], error correcting codes were used for cancelable iris biometrics. A fuzzy commitment method was introduced in [35]. Other schemes have also been introduced to improve the security of iris biometric. See [15, 32, 33, 34, 35, 36] and the references therein for more details.

The pioneering work in the field of cancelable iris biometric was done by Zuo *et al.* [37]. They introduced four non-invertible and revocable transformations for cancelability. While the first two methods utilized random circular shifting and addition, the other two methods added random noise patterns to the iris features to transform them. As noted by the authors, the first two methods gradually reduce the amount of information available for recognition. Since they are essentially linear transformations on the feature vectors, they are sensitive to outliers in the feature vector that arise due to eyelids, eye lashes and specular reflections. They also combine the good and bad quality regions in the iris

image leading to lower performance. The proposed random projections based cancelability algorithm works on each sector of the iris separately, so outliers can only affect the corresponding sectors and not the entire iris vector. Hence, it is more robust to common outliers in iris data when compared to [37].

2.2 Sparse Representation Framework

Following [13], in this section, we briefly describe how to capture the underlying structure in the clean training images using Sparse Representations and utilize it to estimate the class and quality of the inidividual test samples.

Sparse Representations: Suppose that we are given L distinct classes and a set of n training iris images per class. We extract an N-dimensional vector of Gabor features from the iris region of each of these images. Let $\mathbf{D}_k = [\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kj}, \ldots, \mathbf{x}_{kn}]$ be an $N \times n$ matrix of features from the k^{th} class, where $\mathbf{x}_{kj}$ denote the Gabor feature from the j^{th} training image of the k^{th} class. Define a new matrix or dictionary $\mathbf{D}$, as the concatenation of training samples from all the classes as

$$\begin{aligned}\mathbf{D} &= [\mathbf{D}_1, ..., \mathbf{D}_L] \in \mathbb{R}^{N \times (n.L)} \\ &= [\mathbf{x}_{11}, ..., \mathbf{x}_{1n} | \mathbf{x}_{21}, ..., \mathbf{x}_{2n} | ..... | \mathbf{x}_{L1}, ..., \mathbf{x}_{Ln}].\end{aligned}$$

We consider an observation vector $\mathbf{y} \in \mathbb{R}^N$ of unknown class as a linear combination of the training vectors as

$$\mathbf{y} = \sum_{i=1}^{L} \sum_{j=1}^{n} \alpha_{ij} \mathbf{x}_{ij} \tag{2.1}$$

with coefficients $\alpha_{ij} \in \mathbb{R}$. The above equation can be written more compactly as

$$\mathbf{y} = \mathbf{D}\alpha, \tag{2.2}$$

where $\alpha = [\alpha_{11}, ..., \alpha_{1n} | \alpha_{21}, ..., \alpha_{2n} | ..... | \alpha_{L1}, ..., \alpha_{Ln}]^T$ and $.^T$ denotes the transposition operation. We assume that given sufficient training samples of the k^{th} class, $\mathbf{D}_k$, any new test image $\mathbf{y} \in \mathbb{R}^N$ that belongs to the same class will lie approximately in the linear span of the training samples from the class k. This implies that most of the coefficients not associated with class k in (2.2) will be close to zero. Hence, α will be a sparse vector.

Sparse Recovery: In order to represent an observed vector $\mathbf{y} \in \mathbb{R}^N$ as a sparse vector α, one needs to solve the system of linear equations (2.2). Typically $L.n \gg N$ and hence the system of linear equations (2.2) is under-determined and has no unique solution. It has been shown that if α is sparse enough and $\mathbf{D}$ satisfies certain properties, then the sparsest α can be recovered by solving the following optimization problem [38] [39] [40]

$$\hat{\alpha} = \arg\min_{\alpha'} \| \alpha' \|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{D}\alpha', \tag{2.3}$$

where $\| x \|_1 = \sum_i |(x_i)|$. This problem is often known as Basis Pursuit (BP) and can be solved in polynomial time [41][1]. When noisy observations are given, Basis Pursuit DeNoising (BPDN) can be used to approximate α

$$\hat{\alpha} = \arg\min_{\alpha'} \| \alpha' \|_1 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{D}\alpha'\|_2 \leq \varepsilon, \tag{2.4}$$

where we have assumed that the observations are of the following form

$$\mathbf{y} = \mathbf{D}\alpha + \eta \tag{2.5}$$

[1]Note that the ℓ_1 norm is an approximation of the the ℓ_0 norm. The approximation is necessary because the optimization problem in (2.3) with the ℓ_0 norm (which seeks the sparsest α) is NP-hard and computationally difficult to solve.

with $\| \eta \|_2 \leq \varepsilon$.

Sparse Recognition: Given an observation vector $\mathbf{y}$ from one of the L classes in the training set, we compute its coefficients $\hat{\alpha}$ by solving either (2.3) or (2.4). We perform classification based on the fact that high values of the coefficients $\hat{\alpha}$ will be associated with the columns of $\mathbf{D}$ from a single class. We do this by comparing how well the different parts of the estimated coefficients, $\hat{\alpha}$, represent $\mathbf{y}$. The minimum of the representation error or the residual error is then used to identify the correct class. The residual error of class k is calculated by keeping the coefficients associated with that class and setting the coefficients not associated with class k to zero. This can be done by introducing a characteristic function, $\Pi_k : \mathbb{R}^n \to \mathbb{R}^n$, that selects the coefficients associated with the k^{th} class as follows

$$r_k(\mathbf{y}) = \|\mathbf{y} - \mathbf{D}\Pi_k(\hat{\alpha})\|_2. \tag{2.6}$$

Here the vector Π_k has value one at locations corresponding to the class k and zero for other entries. The class, d, which is associated with an observed vector, is then declared as the one that produces the smallest approximation error

$$d = \arg\min_k r_k(\mathbf{y}). \tag{2.7}$$

We now summarize the sparse recognition algorithm as follows:

Given a matrix of training samples $\mathbf{D} \in \mathbb{R}^{N \times (n.L)}$ for L classes and a test sample $\mathbf{y} \in \mathbb{R}^N$:

1. Solve the BP (2.3) or BPDN (2.4) problem.
2. Compute the residual using (2.6).
3. Identify $\mathbf{y}$ using (2.7).

Image quality measure: For classification, it is important to be able to detect and then

reject the test samples of poor quality. To decide whether a given test sample has good quality, we use the notion of Sparsity Concentration Index (SCI) proposed in [13]. The SCI of a coefficient vector $\alpha \in \mathbb{R}^{(L.n)}$ is defined as

$$SCI(\alpha) = \frac{\frac{L.\max \|\Pi_i(\alpha)\|_1}{\|\alpha\|_1} - 1}{L-1}. \tag{2.8}$$

SCI takes values between 0 and 1. SCI values close to 1 correspond to the case where the test image can be approximately represented by using only images from a single class. The test vector has enough discriminating features of its class, so has high quality. If SCI $= 0$ then the coefficients are spread evenly across all classes. So the test vector is not similar to any of the classes and has of poor quality. A threshold can be chosen to reject the iris images with poor quality. For instance, a test image can be rejected if $SCI(\hat{\alpha}) < \lambda$ and otherwise accepted as valid, where λ is some chosen threshold between 0 and 1.

2.3 Bayesian Fusion based Image Selection and Recognition

Different regions of the iris have different qualities [20]. So instead of recognizing the entire iris image directly, we recognize the different regions separately and combine the results depending on the quality of the region. This reduces the computational complexity of the above method as the size of the dictionary is greatly reduced, and the recognition of the different regions can be done in parallel. Also, since occlusions affect only local regions on the iris which can only lower the quality of certain regions, the robustness of the recognition algorithm to occlusion due to eyelids and eye lashes is improved. A direct way of doing this would be to recognize the sectors separately and combine the results by voting [42]. This, however, does not account for the fact that different regions are

recognized with different confidences. In what follows, we propose a score level fusion approach for recognition where we combine the recognition results of different sectors based on the recognition confidence using the corresponding SCI values. Fig. 2 illustrates the different steps involved in the proposed approach.

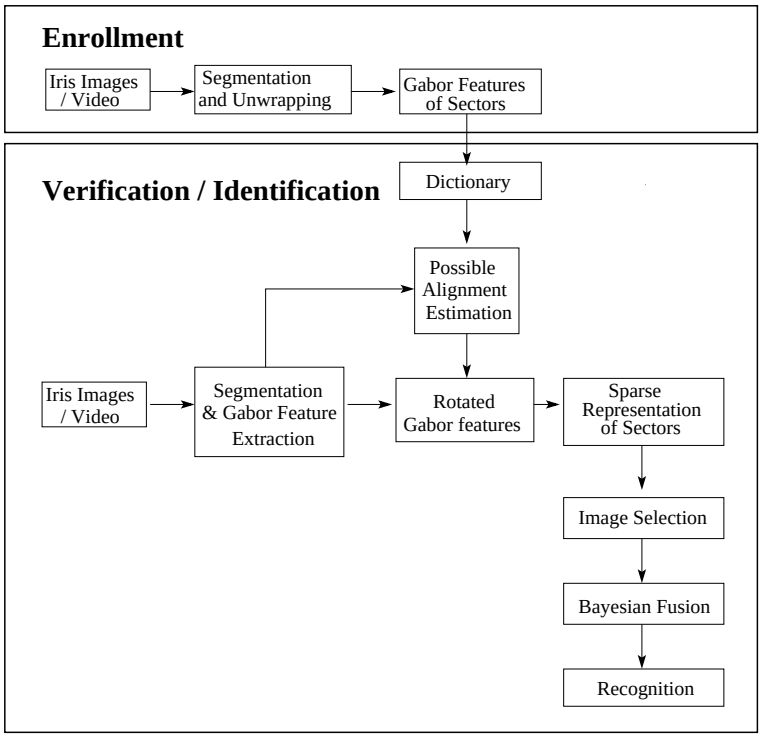


Figure 2.2: A block diagram illustrating the Bayesian Fusion based image selection and recognition.

Consider the iris recognition problem with L distinct classes. Let $\mathbf{C} = \{c_1, c_2, \ldots, c_L\}$ be the class labels. Let $\mathbf{y}$ be the test vector whose identity is to be determined. Let us divide the vector $\mathbf{y}$ into $\hat{M}$ non-overlapping regions, each called a sector. Each of the sectors is individually solved using the sparse representation-based recognition algorithm discussed in section 2.2. The sectors with SCI values below the threshold are rejected. Let M be the number of sectors retained, where $M \leq \hat{M}$. Let $d_1, d_2, \ldots, d_M$ be the class labels of the

retained sectors. Ideally, if the data is noise free, all the returned labels will be equal to the true label c. That is,

$$d_1 = d_2 = \ldots = d_M = c.$$

However, in the presence of noise in the training and test iris images, the returned labels will not necessarily be the same. Let $\mathbb{P}(d_i|c)$ be the probability of the i^{th} sector returns a label d_i when the true class is c. It is reasonable to assume that the probability of the recognition system returning the true label c is high. But given the noise in the iris images, all the classes other than c will still have a low probability of being identified as the true class. SCI is a measure of the confidence in recognition, so the higher the SCI value, the higher the probability that the true class will be the same as the class suggested by the recognition system. So a reasonable model for the likelihood is

$$\mathbb{P}(d_i|c) = \begin{cases} \frac{t_1^{SCI(d_i)}}{t_1^{SCI(d_i)} + (L-1).t_2^{SCI(d_i)}} & if \;\; d_i = c, \\ \frac{t_2^{SCI(d_i)}}{t_1^{SCI(d_i)} + (L-1).t_2^{SCI(d_i)}} & if \;\; d_i \neq c \end{cases} \tag{2.9}$$

where t_1 and t_2 are positive constants such that

$$t_1 > t_2 > 1$$

The numerator gives a higher probability value to the correct class, and the denominator is a normalizing constant. The condition (2.3) ensures that the probability of the true class increases monotonically with the SCI value of the sector. Thus, this likelihood function satisfies the two constraints mentioned above.

The maximum aposteriori estimate (MAP) of the class label given the noisy individual sector labels is given by

$$\tilde{c} = \arg\max_{c \in \mathbf{C}} \mathbb{P}(c|d_1, d_2, \ldots, d_M) \tag{2.10}$$

Assuming the prior probabilities of the classes are uniform, we obtain

$$\tilde{c} = \arg\max_{c \in \mathbf{C}} \mathbb{P}(d_1, d_2, \ldots, d_M | c)$$

Conditioned on the true class, the uncertainty in the class labels is only due to the noise in the different sectors, which are assumed to be independent of each other. So

$$\begin{aligned} \tilde{c} &= \arg\max_{c \in \mathbf{C}} \prod_{j=1}^{M} \mathbb{P}(d_j | c) \\ &= \arg\max_{c \in \mathbf{C}} t_1^{\sum_{j=1}^{M} SCI(d_j).\delta(d_j = c)} . t_2^{\sum_{j=1}^{M} SCI(d_j).\delta(d_j \neq c)} \end{aligned} \tag{2.11}$$

where $\delta(.)$ is the Kronecker delta function. Since $t_1 > t_2$, the solution to (2.11) is same as

$$\tilde{c} = \arg\max_{c \in \mathbf{C}} \sum_{j=1}^{M} SCI(d_j).\delta(d_j = c) \tag{2.12}$$

Let us define the Cumulative SCI (CSCI) of a class c_l as

$$CSCI(c_l) = \frac{\sum_{j=1}^{M} SCI(d_j).\delta(d_j = c_l)}{\sum_{j=1}^{M} SCI(d_j)} \tag{2.13}$$

So

$$\tilde{c} = \arg\max_{c \in \mathbf{C}} CSCI(c) \tag{2.14}$$

CSCI of a class is the sum of the SCI values of all the sectors identified by the classifier as belonging to that class. Therefore, the optimal estimate is the class having the highest CSCI.

2.4 Iris Recognition from video

In this section, we illustrate how our method can be extended to perform recognition from iris videos. Let $Y = \{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^J\}$ be the J vectorized frames in the test video. As before,

each frame is divided into $\hat{M}$ sectors and recognized separately by the sparse recognition algorithm. Let M_i be the number of sectors retained by the selection scheme in the i^{th} frame. Let $\mathbf{y}_j^i$ be the j^{th} retained sector in the i^{th} frame. Using a derivation similar to the one given in Section 2.3, we can derive the MAP estimate as

$$\tilde{c} = \arg\max_{c \in \mathbf{C}} \sum_{i=1}^{J} \sum_{j=1}^{M_i} SCI(d_j^i).\delta(c = d_j^i) \tag{2.15}$$

where d_j^i is the class label assigned by the classifier to $\mathbf{y}_j^i$. (2.15) can be alternatively written as

$$\tilde{c} = \arg\max_{c \in \mathbf{C}} CSCI(c) \tag{2.16}$$

where CSCI of a class c_l is given by

$$CSCI(c_l) = \frac{\sum_{i=1}^{J} \sum_{j=1}^{M_i} SCI(d_j^i).\delta(d_j^i = c_l)}{\sum_{i=1}^{J} \sum_{j=1}^{M_i} SCI(d_j^i)}. \tag{2.17}$$

As before, the MAP estimate consists of selecting the class having the highest cumulative SCI value, with the difference that the sectors of all the frames in the test video will be used while computing the CSCI of each class. Note that unlike existing feature level and score level fusion methods available for iris recognition, the CSCI incorporates the quality of the frames into the matching score. Hence, when the frames in the video suffer from acquisition artifacts like blurring, occlusion and segmentation errors, the proposed matching score gives higher weights to the good frames, at the same time, suppressing the evidence from the poorly acquired regions in the video.

The different modes of operation of the proposed algorithm are illustrated in Fig. 3. Both the probe and the gallery can be separate iris images or iris videos. The iris images are segmented and unwrapped to form rectangular images. The Gabor features of the different

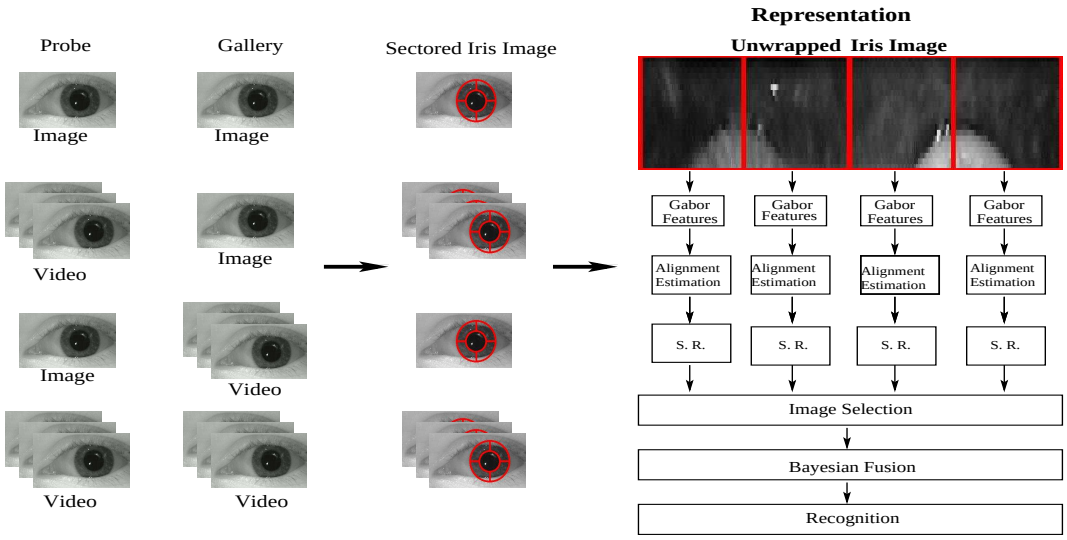


Figure 2.3: A block diagram illustrating the different modes of operation of the proposed algorithm. Both the probe and the gallery can be individual iris images or iris video. Here, S.R. stands for Sparse Representation.

sectors are computed, and sparse representation-based recognition algorithm described in section 2.2 is used to select the good iris images. The good sectors are separately recognized and combined to obtain the class of probe image or video as described above.

2.5 Handling Alignment

Due to rotation of the head with respect to the camera, the captured test iris image may be rotated with respect to the training images. To obtain a good recognition performance, it is important to align the test images before recognition. In this section, we propose a two stage approach for iris feature alignment. In the first stage, we estimate the best K alignments for each test vector using matched filters and then obtain an alignment invariant score function, based on the Bayesian fusion framework introduced above.

2.5.1 Matched Filter Based Alignment Estimation

Let $\mathbf{y}$ be the test vector to be recognized. Let $\hat{A}$ be the number of possible alignments of the test vector. A matched filter is designed for each alignment, whose impulse response is equal to the corresponding shifted version of $\mathbf{y}$. Let $\mathbf{h}_i$ be the impulse response of the i^{th} matched filter, and $\mathbf{H}$ be the set of all possible impulse responses.

$$\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{\hat{A}}\} \tag{2.18}$$

Let e_{ijk} be the sum of squared error between i^{th} matched filter impulse response and j^{th} training image of the k^{th} class.

$$e_{ijk} = \|\mathbf{h}_i - \mathbf{x}_{kj}\|_2^2 \tag{2.19}$$

The alignment error associated with the i^{th} alignment is computed as

$$e_i = \min_{k=1,2,\ldots L, j=1,2,\ldots n} e_{ijk} \tag{2.20}$$

The top K alignments are selected as the ones producing the least alignment error e_i.

2.5.2 Score Estimation Robust to Alignment Errors

From each test vector $\mathbf{y}$, we can generate K new test vectors by shifting it according to the corresponding alignments obtained from the method described above. So instead of the J original frames in the video, we now have JK frames. Using arguments similar to the ones in the previous section, we can obtain the CSCI of the l^{th} class c_l as

$$CSCI(c_l) = \frac{\sum_{i=1}^{JK} \sum_{j=1}^{M_i} SCI(d_j^i).\delta(d_j^i = c_l)}{\sum_{i=1}^{JK} \sum_{j=1}^{M_i} SCI(d_j^i)}. \tag{2.21}$$

where M_i are the number of sectors retained in the i^{th} frame. The MAP estimate of the output class is the one with the highest CSCI value. Note that this score estimation handles

the global alignment errors and not the local deformations in the iris pattern. Since our method weighs different sectors based on their quality, sectors having significant local deformations will not have high influence on the final CSCI value due to their lower quality.

2.6 Secure Iris Biometric

For a biometric system to be deployed successfully in a practical application, ensuring security and privacy of the users is essential. In this section, we propose two cancelable methods to improve security of our recognition system.

2.6.1 Cancelability through Random Projections

The idea of using Random Projections (RP) for cancelability in biometrics has been previously introduced in [36], [43], [44]. In [36] and [43], RPs of discriminative features were used for cancelability in face biometrics. RPs on different regions of the iris were applied for cancelability in [44]. In what follows, we show how RPs can be extended into the sparse representation-based approach for ensuring cancelability.

Let Φ be an $m \times N$ random matrix with $m \leq N$ such that each entry $\phi_{i,j}$ of Φ is an independent realization of q, where q is a random variable on a probability measure space (Ω, ρ). Consider the following observations:

$$\mathbf{a} \doteq \Phi\mathbf{y} = \Phi\mathbf{D}\alpha + \eta', \tag{2.22}$$

where $\eta' = \Phi\eta$ with $\| \eta' \|_2 \leq \varepsilon'$. $\mathbf{a}$ can be thought of as a transformed version of the

biometric **y**. One must recover the coefficients α to apply the sparse recognition method explained in section 2.2. As m is smaller than N, the system of equations (2.22) is underdetermined and a unique solution of α is not available. Given the sparsity of α, one can approximate α by solving the BPDN problem. It has been shown that for sufficiently sparse α and under certain conditions on $\Phi\mathbf{D}$, the solution to the following optimization problem will approximate the sparsest near-solution of (2.22) [45]

$$\hat{\alpha} = \arg\min_{\alpha'} \| \alpha' \|_1 \quad \text{s. t. } \|\mathbf{a} - \Phi\mathbf{D}\alpha'\|_2 \leq \varepsilon'. \tag{2.23}$$

One sufficient condition for (2.23) to stably approximate the sparsest solution of (2.22), is the Restricted Isometry Property (RIP)[46, 40]. A matrix $\Phi\mathbf{D}$ satisfies the RIP of order K with constants $\delta_K \in (0,1)$ if

$$(1-\delta_K) \| v \|_2^2 \leq \| \Phi\mathbf{D}v \|_2^2 \leq (1+\delta_K) \| v \|_2^2 \tag{2.24}$$

for any v such that $\| v \|_0 \leq K$. When RIP holds, $\Phi\mathbf{D}$ approximately preserves the Euclidean length of K-sparse vectors. When $\mathbf{D}$ is a deterministic dictionary and Φ is a random matrix, the following theorem on the RIP of $\Phi\mathbf{D}$ can be stated.

Theorem 1. *([45]) Let $\mathbf{D} \in \mathbb{R}^{N\times(n.L)}$ be a deterministic dictionary with restricted isometry constant $\delta_K(\mathbf{D}), K \in \mathbb{N}$. Let $\Phi \in \mathbb{R}^{m\times N}$ be a random matrix satisfying*

$$P\left(|\|\Phi v\|^2 - \|v\|^2| \geq \varsigma \|v\|^2\right) \leq 2e^{-c\frac{n}{2}\varsigma^2}, \quad \varsigma \in (0, \frac{1}{3}) \tag{2.25}$$

for all $v \in \mathbb{R}^N$ and some constant $c > 0$ and assume

$$m \geq C\delta^{-2}\left(K\log((n.L)/K) + \log(2e(1+12/\delta)) + t\right) \tag{2.26}$$

for some $\delta \in (0,1)$ and $t > 0$. Then, with probability at least $1 - e^{-t}$, the matrix $\mathbf{\Phi D}$ has restricted isometry constant

$$\delta_K(\mathbf{\Phi D}) \leq \delta_K(\mathbf{D}) + \delta(1 + \delta_K(\mathbf{D})). \tag{2.27}$$

The constant satisfies $C \leq 9/c$.

The above theorem establishes how the isometry constants of **D** are affected by multiplication with a random matrix $\mathbf{\Phi}$. Note that one still needs to check the isometry constants for the dictionary **D** to use this result. However, for a given dictionary, **D**, it is difficult to prove that **D** satisfies a RIP. One can alleviate this problem by using the phase transition diagrams [47], [48]. See section VII-A for more details.

The following are some matrices that satisfy (2.25) and hence can be used as random projections for cancelability.

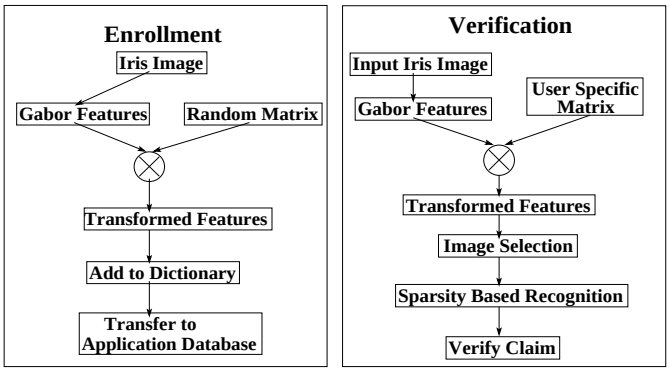
- $m \times N$ random matrices $\mathbf{\Phi}$ whose entries $\phi_{i,j}$ are independent realizations of Gaussian random variables $\phi_{i,j} \sim N\left(0, \frac{1}{m}\right)$.
- Independent realizations of ± 1 Bernoulli random variables

$$\phi_{i,j} \doteq \begin{cases} +1/\sqrt{m}, & \text{with probability } \frac{1}{2} \\ -1/\sqrt{m}, & \text{with probability } \frac{1}{2}. \end{cases}$$

- Independent realizations of related distributions such as

$$\phi_{i,j} \doteq \begin{cases} +\sqrt{3/m}, & \text{with probability } \frac{1}{6} \\ 0, & \text{with probability } \frac{2}{3} \\ -\sqrt{3/m}, & \text{with probability } \frac{1}{6}. \end{cases}$$

- Multiplication of any $m \times N$ random matrix $\mathbf{\Phi}$ with a deterministic orthogonal $N \times N$ matrix $\tilde{\mathbf{D}}$, i.e. $\mathbf{\Phi}\tilde{\mathbf{D}}$.



(a) (b)

Figure 2.4: Block Diagram of the Random Projections based cancelable system.

Note that RPs meet the various constraints required for cancelability, mentioned in Section 2.1. By using different RP matrices, we can issue different templates for different applications. If a transformed pattern is compromised, we can reissue a new pattern by applying a new random projection to the iris vector. The RIP properties together with the sparsity of α ensure that the recognition performance is preserved. In the application database, only the transformed dictionary $\mathbf{\Phi D}$ is stored. If a hacker illegally obtains the transformed dictionary $\mathbf{\Phi D}$ and the transformed iris patterns of the user, $\mathbf{a}$, he or she will have access to the person's identity. However, it is extremely difficult to obtain the matrix $\mathbf{D}$ from $\mathbf{\Phi D}$, and without $\mathbf{D}$ one cannot obtain the original iris patterns $\mathbf{y}$. Hence, our cancelable scheme is non-invertible as it is not possible to obtain the original iris patterns from the transformed patterns. Furthermore, since our method is based on pseudo-random number generation, we only consider the state space corresponding to the value taken by the seed of the random number generator. Hence, instead of storing the entire matrix, one only needs to store the seed used to generate the RP matrix.

2.6.2 Cancelability through Random Permutations of dictionary columns

As explained in section 2.2, when the iris image has good quality, only the training images corresponding to the correct class will have high coefficients. If the training images of different classes are randomly arranged as columns of the dictionary, both the dictionary and the order of the training images are required for correct recognition. In this section, we explain how this idea can enhance the security of our iris recognition system.

When a new user is enrolled, his training images are divided into sectors and placed at random locations in the dictionary. In Fig. 2.5, we show the dictionary for a trivial example of four users. Note that the different sectors of each training image of the user are kept at different random locations in the dictionary. Without prior knowledge of these locations, it is impossible to perform recognition.

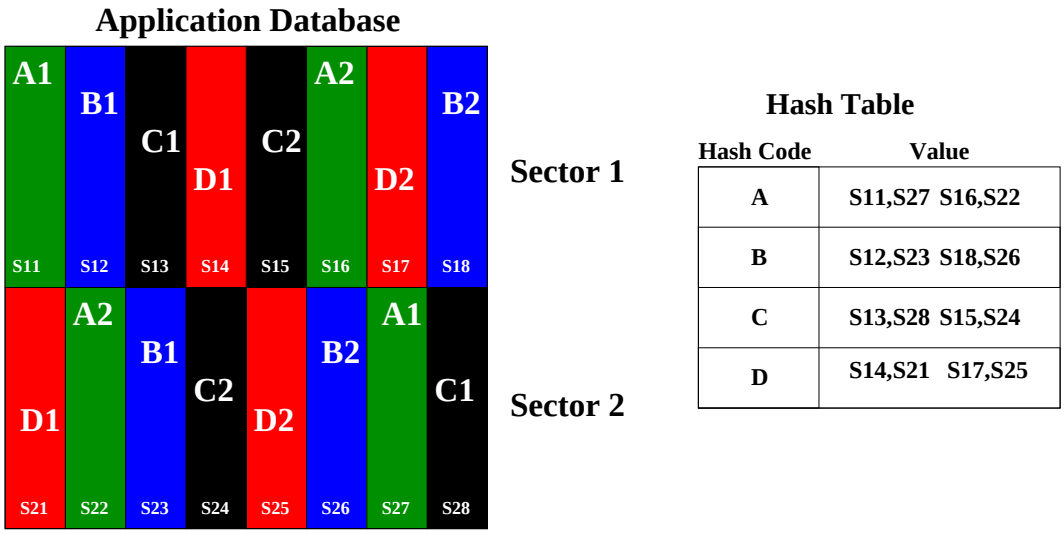


Figure 2.5: Sample Dictionary and hash table for a four user example. The four users A, B, C and D are indicated by colors green, blue, black and red, respectively. A1 and A2 are the two training images corresponding to the first user. $Si\,j$ denote that the j^{th} location and the i^{th} sector. $D1$ at $S14$ means that the first sector of the user D is at location $S14$.

An array indicating the column numbers of the training images of the correct class is generated for each user. This array is stored in a hash table, and the corresponding hash code is given to the user during enrollment. During verification, the system acquires the iris image of the person and extracts the features. For each sector of the iris vector, the sparse coefficients are obtained using this shuffled dictionary, as explained in section 2.2. The user also has to present the hash code to the system. Using the hash code, the indices of training images are obtained from the hash table and the coefficients belonging to different classes are grouped. Then, SCI is computed and used to retain or reject the images. If the image is retained, the CSCI values of the different classes are computed and the class having the lowest CSCI value is assigned as the class label of the user, as explained in section 2.3. A block diagram of the security scheme is presented in Fig. 2.6

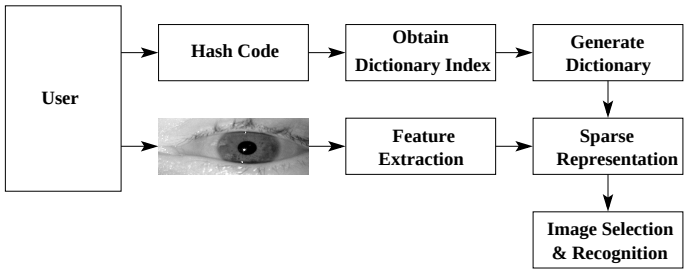


Figure 2.6: Block Diagram of the proposed cancelability scheme using random permutations.

If the hash code presented is incorrect, then the obtained indices of the training images for each class will be wrong. So the coefficients will be grouped in a wrong way, and all the classes will have similar energy leading to a low SCI value and the subsequent rejection of the image. Even if by chance, one of the classes happened to have high energy and the image is retained, the probability of that class being the correct class is very low ($\frac{1}{N}$).

Thus, with high probability, the user will not be verified. Hence, if a hacker illegally acquires the iris patterns of a legitimate user, without having the hash code, he or she will not be able to access the system. Also, even if the hacker obtains the iris dictionary stored in the application database, the iris patterns of the user cannot be accessed without knowing the correct hash codes, because different sectors of an iris patterns reside at different random locations. If the hash code is compromised, the dictionary indices of the user can then be stored at a new location, and a new hash code can be issued to the user. Also, different applications can have different dictionaries. Thus, the user will have a different hash code for each application, preventing cross matching.

It should be noted that the additional security and privacy introduced by these techniques come at the expense of storing additional seed values. In applications requiring higher security, this can be stored with the user, so that a hacker will not get the original templates even if he gets hold of the cancelable patterns in the template database. For applications with greater emphasis on usability, the seed can be stored securely in the template database, so that the user will not have to carry it.

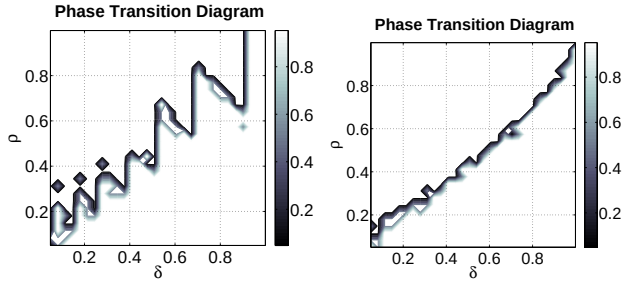
2.7 Results and Discussion

In the following subsections, we present iris image selection, recognition and cancelability results on the ICE2005 dataset [1], ND-IRIS-0405 (ND) dataset [49] and the MBGC videos [50]. The ND dataset is a superset of the ICE2005 and ICE2006 iris datasets. It contains about sixty five thousand iris images belonging to three hundred and fifty six persons, with a wide variety of distortions, facilitating the testing and performance evalu-

ation of our algorithm. In all of our experiments, we employed a highly efficient algorithm suitable for large scale applications, known as the Spectral Projected Gradient (SPGL1) algorithm [51], to solve the BP and BPDN problems.

2.7.1 Empirical verification of ℓ_0/ℓ_1 equivalence

Our sparse recognition algorithm's performance depends on certain conditions on the dictionary such as incoherence and RIP. However, as stated earlier, it is very difficult to prove any general claim that **D**, **GD**, **ΦD**, or **ΦGD** satisfies a RIP or an incoherence property. To address this, one can use the phase transition diagrams [47]. A phase transition diagram provides a way of checking ℓ_0/ℓ_1 equivalence, indicating how sparsity and indeterminacy affect the success of ℓ_1 minimization [47, 48].



(a) (b)

Figure 2.7: Phase transition diagrams corresponding to the case when the dictionary is (a) **GD** and (b) **ΦGD**, where **G** is the Gabor transformation matrix and **Φ** is the random projection matrix for cancelability. In both figures, we observe a phase transition from lower region where the ℓ_0/ℓ_1 equivalence holds, to the upper region, where one must use combinatorial search to recover the sparsest solution.

Let $\delta = \frac{M}{N}$ be a measure of undersampling factor, and $\rho = \frac{K}{M}$ be a measure of sparsity.

A plot of the pairing of the variables δ and ρ describes a two-dimensional phase space $(\delta, \rho) \in [0,1]^2$. The values of δ and ρ ranged through 40 equispaced points in the interval $[0,1]$ and $N = 800$. At each point on the grid, we recorded the mean number of coordinates at which original and reconstruction differed by more than 10^{-3}, averaged over 20 independent realizations (see [47, 48] for more details).

In Fig. 2.7 (a) and (b), we show the phase transition diagrams corresponding to the case when the dictionary is **GD** and **ΦGD**, respectively. Here, **G** is the Gabor transformation matrix and Φ is an $m \times N$ matrix whose entries $\phi_{i,j}$ are independent realizations of Gaussian random variables $\phi_{i,j} \sim N\left(0, \frac{1}{m}\right)$. For each value of δ, the values of ρ below the curve, are the ones where the ℓ_0/ℓ_1 equivalence holds. As can be observed, for most values of δ, there is atleast one value of ρ below the curve, satisfying the equivalence. So the vector α can be recovered if it is sparse enough and enough measurements are taken.

2.7.2 Image Selection and Recognition

In this section, we evaluate our selection and recognition algorithms on ND and ICE2005 datasets. To illustrate the robustness of our algorithm to occlusion due to eyelids and eyelashes, we perform only a simple iris segmentation scheme, detecting just the pupil and iris boundaries and not the eyelids and eye lashes. We use the publicly available code of Masek *et al.* [52] for detecting these boundaries.

2.7.2.1 Variation of SCI with common distortions during image acquisition

To study the variation of SCI in the presence of common distortions during image acquisition like occlusion and blur, we simulate them on the clean iris images from the ND dataset.

Description of the Experiment: We selected fifteen clean iris images of the left eye of eighty persons. Twelve such images per person formed the gallery and distortions were simulated on the remaining images to form the probes. We consider seven different levels of distortion for each case, with level one indicating no distortion and level seven indicating maximum distortion. We obtain the dictionary using the gallery images, and evaluate the SCI of the various sectors of the test images.

Fig. 2.8 shows some of the simulated images from the ND dataset. The first column includes images with distortion level one (no distortion). The middle column contains images with distortion level three (moderate distortions). The right most column contain images with distortion level five (high distortion). The first row contains images with blur while the second contains images with occlusion. Images with simulated segmentation error and specular reflections are shown in the third and fourth rows respectively.

Fig. 2.9 (a) illustrates the variation of SCI with the common acquisition distortions. It can be observed that good images have high SCI values whereas the ones with distortion have lower SCI values. So by suitably thresholding the SCI value of the test image, we can remove the bad images before the recognition stage. The relative stability in SCI values with occlusion and specular reflection demonstrates the increased robustness attained by

our algorithm, by separately recognizing the individual sectors and combining the results, as mentioned in section 2.3.

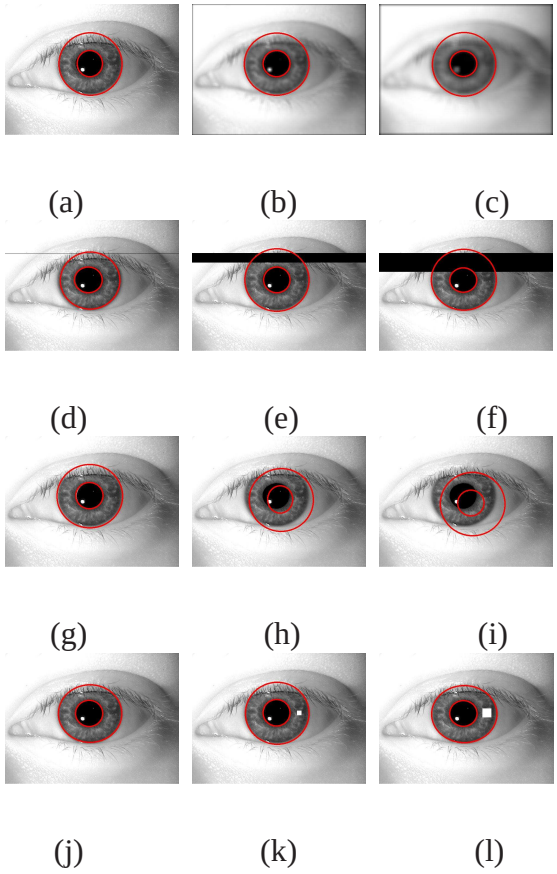
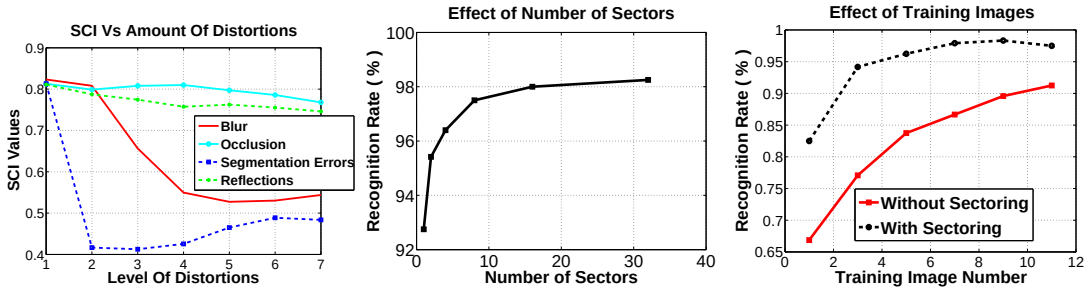


Figure 2.8: Simulated Distortions on the images from the ND dataset. The detected pupil and iris boundaries are indicated as red circles.

2.7.2.2 Image Selection results on the ND dataset

In this section, we illustrate the performance of our image selection algorithm on images from the ND dataset.

Description of the Experiment: We selected the left iris images of eighty subjects that had sufficiently large number of iris images with different distortions like blur, occlusion and segmentation errors. Fifteen clean images per person were hand chosen to form the



(a) (b) (c)

Figure 2.9: (a) Plot of the variation in SCI values with common distortions in iris image acquisition. Note that the SCI falls monotonically with increasing levels of blur and segmentation errors in the iris images. It is also robust to occlusions and specular reflections. (b) Plot of the recognition rate versus the number of sectors. Observe the significant improvement in the results as the number of sectors is improved from one to eight. (c) Plot of the recognition rate versus the number of training images. Note that the recognition rate increases monotonically with the number of training images. Also, sectoring achieves the same recognition rate as the case without sectoring using far fewer training images.

gallery. Up to fifteen images with blur, occlusion and segmentation errors were also selected. As mentioned before, we perform a simple segmentation scheme, retaining possible occlusion due to eyelids and eyelashes in the iris vector. The Gabor features of the iris vector form the input. Our algorithm creates the dictionary, finds the sparse representation for each test vector, evaluates the SCI of the sectors, and rejects the images for which all the sectors have SCI value below a threshold of 0.6.

Measure the selection performance : The quality of the input iris feature vector should be a function of the performance of the recognition algorithm on that sample [12]. An ideal image selection algorithm should retain images, which can be correctly recognized by the recognition algorithm, and reject the ones on which the subsequent recognition algorithm will perform poorly. To measure it, we define the Modified False Positive Rate (MFR) and a Modified Verification Rate (MVR) as follows. Modified False Positive rate is the fraction of the test vectors retained by the image selection algorithm, which are wrongly classified by the subsequent recognition algorithm. Modified Verification Rate is defined as the fraction of the images correctly classified by the recognition algorithm, which are retained by the selection scheme. To obtain these values, we find the CSCI for each test sample and also the class assigned to the samples by our algorithm. We obtain the Receiver Operating Characteristics (ROC) of the image selection algorithm by plotting MVR versus MFR for different values of threshold. Note that this measures the performance of the quality estimation stage and is different from the ROC curve of the recognition algorithm.

$$MFR = \frac{\text{No of Images selected and wrongly classified}}{\text{No of images selected}}$$

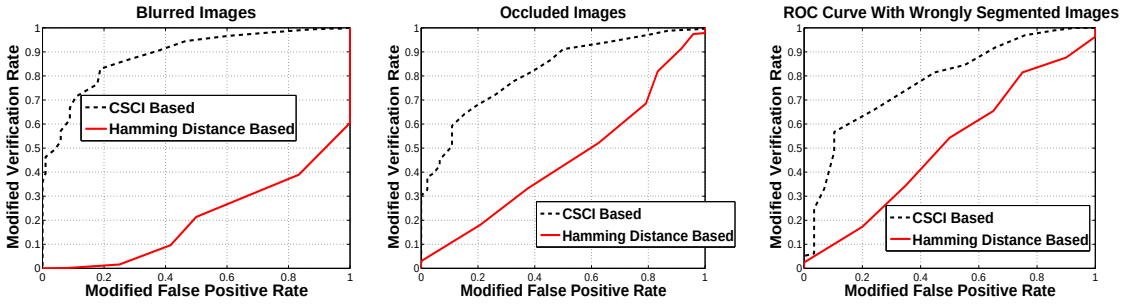
$$MVR = \frac{\text{No of Images selected and correctly classified}}{\text{No of images correctly classified}}$$

Fig. 2.10(a) shows the ROC of our image selection algorithm (black), compared to that using directly the Hamming distance based on the publicly available iris recognition system of Masek *et al.* [52] (red), when the probe images are blurred. Since the data has occlusion, direct application of Masek's algorithm performed poorly. For a fair comparison, we modified the algorithm, recognizing the different sectors of the iris separately and fusing the results through voting. Note that our ROC curve is significantly sharper than that of the Masek's recognition system indicating superior performance.

The effects of occlusion in iris images due to eyelids, eye lashes and specular reflections are illustrated in Fig. 2.10(b). Images with occlusion were obtained for each of the eighty classes under consideration and used as probes. The ROC curve of our algorithm is shown in black and that of Masek's system appears in red. Note that for the same MFR, the proposed image selection scheme has a higher MVR. This indicates that the proposed selection method retains more images that will be correctly classified by the subsequent recognition algorithm and rejects more images that will be wrongly classified by the recognition algorithm.

To study the effects of segmentation error, the gallery images were verified to be well segmented. Up to fifteen images with segmentation errors were chosen for each person under consideration, which formed the probes. Fig. 2.10(c) shows the ROC curves of our method (black) and the Masek's one (red) in case of wrongly segmented images. Again, using our image selection algorithm improves the performance of the system even with wrongly segmented images, a feature lacking in many existing quality estimation

methods.



(a) (b) (c)

Figure 2.10: Comparison of the ROC curves of the proposed image selection algorithm (CSCI Based) and one using Hamming distance as the quality measure(Hamming Distance Based) using clean iris images in the gallery and probe images containing (a) Blurring (b) Occlusions and (c) Segmentation Errors. Note that CSCI based image selection performs significantly better than Hamming distance based selection when the image quality is poor.

2.7.2.3 Recognition Results on images from the ND dataset

In this section, we illustrate the performance of our recognition algorithm on images from the ND dataset.

Performance on clean images - Description of the Experiment: Eighty subjects were selected from the dataset. Fifteen clean images of the left iris were hand selected for each person. Of these fifteen images per person, twelve were randomly selected to form the gallery and the remaining three images per person were used as probes. No image selection is performed because we want to evaluate the performance of the recognition

algorithm separately.

We compare our algorithm to a nearest neighbor based recognition algorithm (NN) that uses the Gabor features and the Masek's implementation. Since we use tough segmentation conditions retaining the eyelids and eye lashes in the iris vector, direct application of NN and Masek's method produced poor results. For a fair comparison, we divided the iris images into different sectors, obtained the results using these methods separately on each sectors and combined the results by voting. We obtained a recognition rate of 99.15% when compared to 98.33% for the NN and 97.5% for the Masek's method.

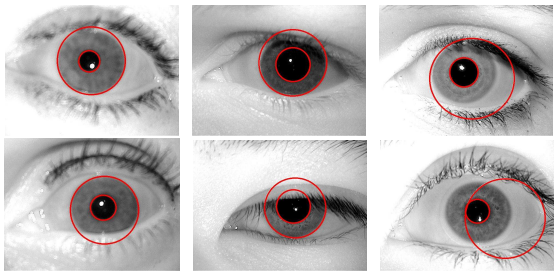
Performance on poorly acquired images - Description of the Experiment - To evaluate the recognition performance of our algorithm on poorly acquired images, we hand picked images with blur, occlusion and segmentation errors as explained in the previous section. Fifteen clean images per person were used to form the gallery. Probes containing each type of distortion were applied separately to the algorithm. We perform image selection followed by recognition. The recognition rates are reported in Table. 2.2.

Table 2.1: Recognition Rate On ND Dataset

Image Quality	NN	Masek's Implementation	Proposed Method
Good	98.33	97.5	99.15
Blurred	95.42	96.01	98.18
Occluded	85.03	89.54	90.44
Seg. Error	78.57	82.09	87.63

In Fig. 2.11, we display the iris images having the least SCI value for the blur, occlusion

and segmentation error experiments performed on the real iris images in the ND dataset as mentioned above. As can be observed, images with low SCI values suffer from high amounts of distortion.



(a) (b) (c)

Figure 2.11: Iris images with low SCI values in the ND dataset. Note that the images in (a), (b) and (c) suffer from high amounts of blur, occlusion and segmentation errors respectively .

2.7.2.4 Recognition Performance on the ICE 2005 Dataset

In this section, we compare the performance of our algorithm with the existing results on the ICE 2005 dataset corresponding to Experiment 1. Experiment 1 has 1425 iris images corresponding to 120 different classes.

Description of the Experiment : We have used ten images per class in the gallery and remaining iris images as the test vectors. We perform segmentation using Masek's code and apply the Gabor features of the segmented iris images to our recognition algorithm. No image selection was performed. We compare our performance with existing results in Table 2.2, where the verification rates are indicated at a false acceptance rate of 0.001. The results of the existing methods are obtained from [53].

Table 2.2: Verification rate at an FAR of 0.001 on the ICE 2005 dataset

Method	Verification Rate (%)
Pelco	96.8
WVU	97.9
CAS 3	97
CAS 1	97.8
CMU	99.5
SAGEM	99.8
Proposed Method	98.13

2.7.2.5 Dependence of recognition rate on the number of sectors

Fig. 2.9 (b) plots the variation of the recognition rates for the proposed method with changes in the number of sectors. As can be observed, the performance of the recognition system improves significantly as the number of sectors is increased from one to eight. Beyond eight, the recognition rate does not increase significantly.

2.7.2.6 Effect of the number of training images on performance

In this section, we study the effect of the number of training images on recognition rate of our algorithm. We vary the number of training images from one per class to eleven per class on the ND dataset. The test images consisting of three iris images per person are used to test each of these cases. The variation of recognition rate is plotted in Fig. 2.9

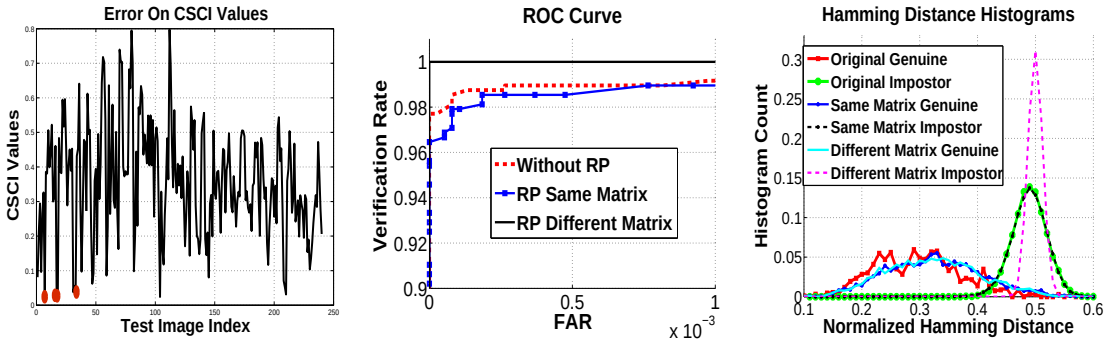
(c) for the case of no sectoring and sectoring with eight sectors respectively. As can be observed, recognition performance increases with the number of training images. This is hardly surprising as our assumption that the training images span the space of testing images becomes more valid as the number of training images increases. In unconstrained iris recognition systems which we are interested in, this is not a bottle neck because we can obtain a significant number of iris images from the incoming iris video. Also, sectoring achieves the same recognition rate as the non-sectoring case with a much lower number of training images.

2.7.2.7 CSCI as a measure of confidence in recognition

We have empirically observed that the higher the CSCI value for the test image, the higher the probability that it is correctly classified. This is illustrated in Fig. 2.12 (a). This observation is expected as high CSCI means that the reconstructed vector in most of the sectors will be sparse. If the training images span the space of possible testing images, the training images of the correct class will have high coefficients. So the only possible sparse vector is the one in which the correct class has high coefficients and others have zero coefficients, which will be correctly classified by our algorithm.

2.7.3 Cancelability Results using Random Projections

We present cancelability results on the clean images from the ND dataset obtained as explained in Section 2.7.2.3. The iris region obtained after segmentation was unwrapped into a rectangular image of size 10×80. The real parts of the Gabor features were ob-



(a) (b) (c)

Figure 2.12: (a) Plot of the CSCI values of test images for a random trial on the ND dataset. Red dots indicate the wrongly classified images. Observe that the wrongly classified images have low CSCI values and hence the corresponding vectors are not sparse. (b) ROC characteristics for the ND dataset. The Same Matrix performance is close to the performance without cancelability . Using different matrices for each class gives better performance. (c) Comparison of the distribution of the Genuine and Impostor normalized Hamming distances for the original and transformed patterns.

tained and concatenated to form an iris vector of length 800. We used the random Gaussian matrix in our experiments, though other random matrices mentioned in Section 2.6.1 also gave similar results. In [44], it was shown that separate application of the random projections performed better when compared to the application of a single random projection on the entire iris vector. So we vectorized the real part of the Gabor features of each sector of the iris image, applied the random projections, and then concatenated the random projected vectors to obtain our cancelable iris biometric. We applied either the same random Gaussian matrix for all the users or different random matrices for different users to obtain the RP "Same Matrix" and "Different Matrix" vectors, respectively. Having obtained the random vectors from the Gabor features of the iris image, we performed the sparsity-based recognition algorithm described in Section 2.2. We present the Receiver Operating Characteristic (ROC) curves and the Hamming distance distributions in the subsections below.

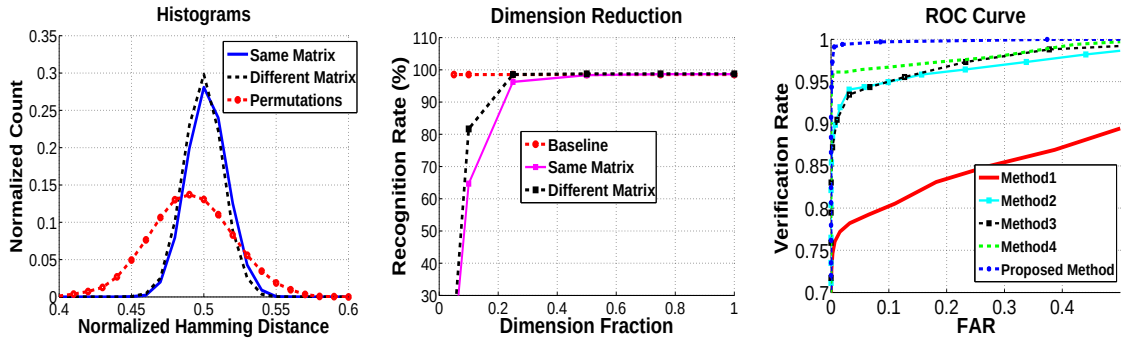
2.7.3.1 Recognition Performance

Fig. 2.12(b) plots the ROC characteristics for the iris images in the ND dataset for the original and transformed iris patterns. As demonstrated, using different matrices for each class performs better than using the same matrix for all classes. In the "Different Matrix" case, we assumed that the user provided the correct matrix assigned to him. So the performance exceeds even the original performance as class specific random projections increases the interclass distance, still retaining the original intra-class distance. In Fig. 2.12 (c), we compare the distribution of the genuine and impostor normalized Hamming dis-

tance for the original and transformed iris patterns. We can observe that the distribution of the genuine Hamming distance remains almost the same after applying the random projections. The original and Same Matrix cases have similar impostor Hamming distance distributions. However the Different Matrix case has an impostor distribution that is more peaked and farther from the genuine distribution, indicating superior performance.

2.7.3.2 Normalized Hamming distance comparison between the original and the transformed patterns

In this section, we quantify the similarity between the original and the random projected iris vectors. From the original and transformed iris vectors, iris codes are computed by allocating two bits for each Gabor value. The first bit is assigned one if the real part of the Gabor feature is positive and zero otherwise. The second bit is assigned a value of one or zero in a similar manner based on the imaginary part of the Gabor feature. The normalized Hamming distance between the iris codes is used as the measure of similarity. In Fig. 2.13(a), we plot the normalized Hamming distance between the iris codes of the original and the transformed iris vectors for the “Same Matrix” and “Different Matrix” cases, respectively. Ideally we want the two iris codes to be independent, hence the normalized Hamming distance should be 0.5. The figure shows that the histogram of the Hamming distance peaks at 0.5, empirically verifying that the random projected iris vectors are significantly different from the originals ones. Hence it is not possible to extract the original iris codes from the transformed version, thereby proving the non-invertibility property of our transformation.



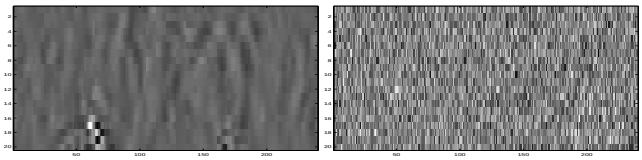
(a) (b) (c)

Figure 2.13: (a) Plot of the histograms of the Normalized Hamming Distance between the original and transformed vectors. Note that the histogram peaks around 0.5 indicating that the original and transformed iris codes are significantly different. (b) Plot of the recognition rate with dimension reductions for the ND dataset. Note that the performance remains the same up to 30% of the original dimension. (c) ROC plots for video based iris recognition. Method 1 treats each frame in the video as a different probe. Method 2 averages all the frames in the probe video. Methods 3 and 4 use the average and minimum of all the pair wise Haming distance between the frames of the probe and gallery videos respectively. The Proposed Method uses CSCI as the matching score. Note that the introduced quality based matching score outperforms the existing fusion schemes, which do not incorporate the quality of the individual frames in the video.

Table 2.5 provides the statistics of the normalized Hamming distance between the original and the transformed iris vectors. As can be seen, the mean of the normalized Hamming distance is very close to 0.5 with a very low standard deviation.

Table 2.3: Statistics Of The Normalized Hamming Distance.

Methods	Mean	Standard Deviation
Without RP	0	0
Same Matrix	0.5002	0.0123
Different Matrix	0.4999	0.013
Dictionary Permutations	0.4913	0.0254



(a) (b)

Figure 2.14: (a) Gabor features of the original iris image. (b) Gabor features of the recovered iris image from the cancelable patterns in the dictionary and a randomly generated projection matrix.

2.7.3.3 Non-Invertibility Analysis of Cancelable Templates using Random Projections

In this section, we consider the recovery of original iris patterns from the cancelable templates, using varying levels of information about the dictionary and the projection matrix Φ. We consider two methods, one based on minimizing the squared error and the

other based on compressive sensing techniques. As before, we consider eighty classes from the ND-IRIS-0405 dataset with fifteen images per class. Twelve images per person for the training set and the remaining for the test vectors. We apply the same random projections Φ for each class with a dimension reduction of 40% to form the cancelable patterns. Hence, we have the $\mathbf{a} = \Phi\mathbf{D}\mathbf{y}$, where $\mathbf{a}$ is the cancelable template and $\mathbf{y}$ is the original iris pattern. We consider two methods for reconstructing the original patterns from cancelable patterns. They are explained below.

1. Least Square solution - From equation (2.22) in the presence of additive noise, the original template can be recovered by minimizing the following squared error.

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}} \|\mathbf{a} - \Phi\mathbf{y}\|_2^2$$

2. Compressive Sensing based solution - Since Φ is a random Gaussian matrix having good RIP, one possible way of reconstructing the iris patterns is by solving the following L1 minimization problem.

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}} \| \mathbf{y} \|_1 \quad \text{s. t.} \quad \|\mathbf{a} - \Phi\mathbf{y}\|_2 \leq \varepsilon'. \tag{2.28}$$

We computed the error in reconstruction of the original patterns and the recognition rate on the reconstructed patterns for different levels of information known about the cancelable template dictionary and the random projection matrix Φ. The results are shown in Table 2.5. As can be observed, the recognition performance is close to chance when either the random matrix or the dictionary entries are not known. Even when the random matrix and the dictionary entries are fully known, the recognition performance on the reconstructed template is significantly lower than that on the original iris templates. This

result empirically verifies that it is difficult to extract significant information about the original iris templates from the cancelable ones.

Table 2.4: Reconstruction Error and Recognition Rate knowing the exact cancelable template and fraction of entries in the projection matrix

Method	Metric %	Fraction Of Correct Values					
		0	.2	.4	.6	.8	1
LS	Recon. Error	50	49	49	49	49	49
	Recog. Rate	2.9	2.08	2.08	.42	.83	.83
CS	Recon. Error	49	46	42	38	32	22
	Recog. Rate	1.67	2.08	3.33	7.92	24.58	59.17

In Fig. 2.14, we display the Gabor features of one of the iris images in the dictionary and the corresponding recovered pattern. As can be observed, the recovered pattern appears as random noise and does not contain any of the information in the original iris pattern.

2.7.3.4 Effect of dimension reduction

In Fig. 2.13(b), we demonstrate the robustness of random projections to reduction in the original dimension of the feature vector. The random projected vectors retain their original performance for up to 30% reduction in the original dimension for both the same and different matrix cases. Dimension reduction further strengthens the non-invertibility of our transformation as there will be infinite possible iris vectors corresponding the reduced dimension random vectors obtained by RP.

Table 2.5: Reconstruction Error and Recognition Rate knowing the exact projection matrix and fraction of entries in the cancelable template

Method	Metric (%)	Fraction Of Correct Values					
		0	.2	.4	.6	.8	1
LS	Recon. Error	49	49	49	49	49	49
	Recog. Rate	1.25	2.08	1.25	.83	1.25	2.5
CS	Recon. Error	49	48	46	43	38	22
	Recog. Rate	1.25	1.67	1.25	1.67	9.17	57.50

2.7.3.5 Comparison with Salting

In Table. 2.6, we present the recognition rates and the corresponding mean Hamming distance for the salting method proposed in [37] for various noise levels. The best recognition rate and the best Hamming distance for the Salting method are 96.6% and 0.494 respectively. For RP Same Matrix case, we obtained a recognition rate of 97% at a Hamming distance of .497. Thus both the recognition performance and security (non-invertibility) are higher for RP when compared to the Salting method.

2.7.4 Cancelability Results using Random Permutations

To evaluate the performance of the proposed cancelable method using dictionary permutations, we consider the three possible scenarios on the clean images from the ND dataset. In the first case, the user provides the iris image and the correct hash code. In this case, the

Table 2.6: Comparison with Salting method. The Recognition Rate(RR) and mean Hamming Distance (HD) are provided for the Salting and SRP methods. The recognition rate obtained using SRP is higher than that of the Salting method. Also SRP gives mean Hamming distance closer to .5 when compared to the Salting method.

Quantity	Salting			Same	Different	Permutations
RR(%)	94.2	96.6	94.0	97	100	100
HD	0	.491	.494	.497	.50	.483

recognition performance was the same as that of the original method on the ND dataset, which is 99.17%. In the second case, the user provides the iris image but a wrong hash code. Here the recognition performance dropped to 2%, which is only slightly better than chance. This is equivalent to the case when a hacker illegally obtains the iris image of a valid user and tries to gain access into the system with a guess about the hash code. The low recognition performance clearly reflects the additional security introduced by the permutations, as a hacker needs to now have not only the iris image but also the hash code of a valid user to gain access. In the third experiment, we found the closeness between the Gabor features of the original iris images and the new feature vectors obtained by permutations of the Gabor features in the dictionary. As before, the normalized Hamming distance between the iris codes obtained from these vectors is used as the measure of similarity. We plot the histogram of the normalized Hamming distance between the original and the randomly permuted iris vectors in Fig. 2.13(a). The mean and standard deviation of the Hamming distance histogram are indicated in the last row of the Table. 2.5. Note that the mean is close to .5, indicating that the permutations differ significantly different

from the original iris images. Even if a hacker can use the dictionary from the application database, he or she will be unable to extract information about the original iris images without knowing the hash code of each user.

2.7.5 Results on Iris Videos

In this section, we present the results on the MBGC videos [50]. Given the thirty classes, we used twenty eight classes that contained atleast five good images in our experiments. We hand picked five clean images from the iris videos in the training set which formed the dictionary. In the test videos, batches of five frames were given as a probe to our algorithm. Using twenty eight available videos and sixty frames from each test video, we could form three hundred and thirty six probes. We did only a basic segmentation of the iris and pupil using the Masek's code, as before. Also, we did not remove the poorly segmented iris images manually before performing the recognition algorithm.

We compare the performance of our algorithm with four other methods. The ROC plots for the different methods are displayed in Fig. 2.13(c). In Method 1, we consider each frame of the video as a different probe. It gave the worst performance, indicating that using multiple frames available in a video can improve the performance. Method 2 averages the intensity of the different iris images. Though it performs well when the images are clean, a single image which is poorly segmented or blurred could affect the entire average. In Methods 3 and 4, all possible pair wise Hamming distances between the video frames of the probe videos and the gallery videos belonging to the same class are computed. Method 3 uses the average of these Hamming distance as the score. In Method 4,

the minimum of the pairwise Hamming distance was used as the score. In the proposed method, the CSCI values were computed for each class for each probe video and the probe video is assigned to the class having the highest CSCI value. For a fair comparison of the proposed quality measure in videos, we did not reject any of the frames. Observe that our method performs better than other methods. One of the reasons for the superior performance could be the fact that we are incorporating the quality of the different frames while computing the CSCI. Frames which are poorly segmented or blurred will have a low SCI value and hence will not affect the score significantly. In all the other methods, the image quality was not effectively incorporated into the matching score, so all frames are treated equally irrespective of their quality.

Chapter 3

Sensor Adaptation in Iris Recognition

As explained in Chapter 2, iris recognition is one of the most popular approaches for non-contact biometric authentication [12]. Over the past decade, sensors for acquiring iris patterns have undergone significant transformations: existing ones have been upgraded and new ones have been developed [54]. These transformations pose new challenges to iris recognition algorithms. Due to the large number of users, possibly in millions, enrollment is expensive and time-consuming. This makes it infeasible to re-enroll users every time a new sensor is deployed. In practice, one often encounters situations where iris images for enrollment and testing are acquired by different sensors.

Recent studies in iris biometrics illustrate that cross-sensor matching, where different sensors are employed for enrollment and testing, often lead to reduced performance [55]. We illustrate this using the LG2200 and LG4000 sensors in Figure 3.1. As can be observed, the receiver operating characteristics (ROC) curve of cross-sensor matching is inferior to that of same-sensor matching. We refer to this performance drop due to the difference in the sensors used for enrollment and testing as the "sensor mismatch" problem in iris recognition, and techniques to alleviate it as "sensor adaptation" methods. While the sensor mismatch problem has been empirically illustrated by [55] and [56], research in algorithms for sensor adaptation specific to iris biometrics has been limited in the literature.

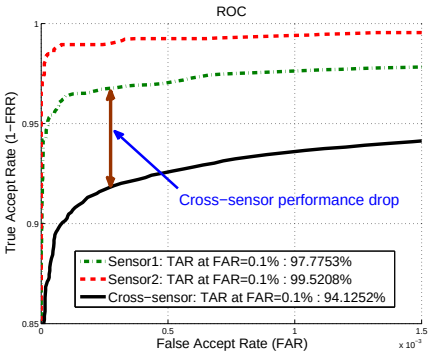


Figure 3.1: ROC curves for the same-sensor and the cross-sensor case, collected under similar acquisition conditions. Observe that the black curve corresponding to cross-sensor matching is significantly lower than the same-sensor matching curves in red and green, indicating the performance drop caused by sensor mismatch.

In this chapter, we first present a novel optimization framework for learning transformations of iris biometrics having the desired properties. These transformations can be concisely represented using kernel functions. The proposed framework is then utilized for sensor adaptation, by constraining the samples from different sensors to behave in a similar manner in the transformed domain. Specifically, we enforce the following constraints on the transformation. In the transformed space, the distances between iris samples belonging to the same class should be small, irrespective of the sensor used for their acquisition. Furthermore, those between samples of different classes should be large. These constraints ensure that the sensor mismatch problem is alleviated, when cross-sensor matching is performed in the transformed domain.

While the original optimization problem is convex and has a global optimum, it needs to be performed every time a test sample is acquired. Hence, it is computationally expensive. By rewriting the optimization problem, an efficient solution is obtained using Bregman

projections. This solution involves estimating the adaptation parameters during the training stage. During testing, the test iris samples are transformed using these parameters. Cross-sensor matching is then performed using the transformed iris samples. Since the learned transformation alleviates the sensor mismatch problem, cross-sensor matching in the transformed domain leads to significant improvements in accuracy.

Contributions: The main contributions of this work are:

1. The proposed method is one of the first comprehensive solution for the sensor mismatch problem in iris biometrics.
2. The introduced solution leads to considerable improvement in cross-sensor matching. It is robust to alignment errors, and can also handle real-valued feature representations.
3. The proposed technique is fast, requiring limited changes to the existing iris recognition pipeline. Hence, it can easily be incorporated into existing iris recognition systems.

Organization of the chapter: The relevant literature in iris recognition and machine learning is described in Section 3.1. In Section 3.2, a similarity measure is developed for iris codes and its properties are analyzed. A general optimization framework for learning kernel functions for iris codes is introduced in Section 3.3. The sensor mismatch problem is formulated as a kernel learning problem in Section 3.4. By reformulating this optimization problem using the similarity measure introduced in Section 3.2, an efficient solution is developed in Section 3.5. The proposed method is evaluated on iris data from multiple sensors in Section 3.6.

3.1 Related Work

3.1.1 Iris Recognition

The main components in an iris recognition system are image acquisition, iris segmentation, feature extraction, and template matching [12]. Please refer to the 'Related Work' section in Chapter 2 for a detailed description of the existing literature on iris recognition.

3.1.2 Iris Acquisition Systems

Iris image acquisition systems differ mainly in the type and location of the illumination they use, the type of sensor, and the presence of additional optical elements [54]. Due to the different design possibilities and significant commercial interests in iris recognition, numerous iris acquisition systems are available, with the potential for many more. Some of the popular systems are LG2200, LG4000, Iris on the Move portal system by Sarnoff, Combined Face And Iris Recognition System (CFAIRs) by Honeywell, HBOX^{TM} system by Global Rainmakers Inc., and Eagle-Eyes^{TM} system by Retica. Interested readers are referred to [54] for a detailed review of these systems.

3.1.3 Sensor Interoperability

Owing to the large number of iris recognition systems currently available and the continuous improvement of existing systems, the inter-operability of iris systems become extremely important. In the past, several works have addressed the problem of biometric interoperability for fingerprint sensors [57] [58], or multibiometric systems [59]. In

iris biometrics, this problem was first investigated by Bowyer *et. al* [55] using two iris sensors. Their work demonstrated that the older of the two sensors provided less desirable match score distributions. Furthermore, the cross-sensor performance was inferior to that of either sensors tested individually. Cross-sensor iris recognition was further explored by Connaughton *et. al* [56], who experimented with three commercially available iris sensors. These methods clearly demonstrate the need for improving the cross-sensor recognition performance.

3.1.4 Kernel Methods in Machine Learning

Since we follow a kernel-based approach for sensor adaptation, a brief introduction to kernel methods in machine learning is provided in this section. Interested readers are referred to [60] for an extensive description of the topic. To capture non-linear relationships, kernel methods project the data into a higher dimensional space and fit linear models in the projected space. Data appear in computation only in the form of inner products, which can be performed without explicit projection into the high dimensional space, using kernel functions. Boser *et. al* [61] introduced kernels into mainstream machine learning literature by combining kernel functions and maximum margin hyperplanes, leading to nonlinear support vector machines (SVM). Kernels have also been used for metric learning [62], domain adaptation [63], and dictionary learning [64]. Specialized kernel functions have been developed for different applications, such as text categorization [65] and scene analysis [66]. Furthermore, kernel functions have also been developed in an optimization framework, where desired properties are enforced by the chosen constraints [67]. This

framework is explained in detail in Section 3.3.

3.2 Similarity Measure

In this section, a similarity measure is introduced for iris samples and its properties are analyzed. This measure will play an important role in developing an efficient sensor adaptation algorithm in Section 3.5.

Notations: We first introduce the notations used in the paper. Let $\mathscr{B}^D = \{0,1\}^D$ be the space of all binary vectors of length D. Let the iris samples available during training be denoted by $\mathscr{L} = \{\theta_1, \theta_2, \ldots, \theta_N\}$. Here, the i^{th} iris sample $\theta_i \in \mathscr{B}^{2D}, \theta_i^T = \left[x_i^T \; m_i^T\right]$, where $x_i \in \mathscr{B}^D$ is the i^{th} iris code and $m_i \in \mathscr{B}^D$, the corresponding mask. Let $y_i \in \{1,2,\ldots N_c\}$ denote the class label of the i^{th} iris sample and $s_i \in \{1,2,\ldots,N_s\}$ denote the sensor from which it was acquired. Here, N denotes the number of training samples, D the dimension of the iris codes, N_c the number of subjects enrolled, and N_s the number of sensors used for acquisition. We denote the j^{th} bit in the i^{th} iris code by $x_i(j)$. $x_i(j)$ is called a "valid" bit if the corresponding masking bit $m_i(j) = 1$. Furthermore, let $\wedge, \oplus$ and $\neg$ denote the logical AND, XOR and NOT operations, respectively.

The normalized Hamming distance $\mathscr{H}(\theta_i, \theta_j)$ between two iris samples θ_i and θ_j is defined as the fraction of the valid bits that disagree [68]. So

$$\mathscr{H}(\theta_i, \theta_j) = \frac{\sum_{l=1}^{D}\{m_i(l) \wedge m_j(l) \wedge (x_i(l) \oplus x_j(l))\}}{\sum_{l=1}^{D}\{m_i(l) \wedge m_j(l)\}}. \tag{3.1}$$

3.2.1 Definitions

Given two iris samples θ_i and θ_j, we define the joint agreement $\eta_A(\theta_i, \theta_j)$ as the number of valid bits that agree between θ_i and θ_j. Similarly, the joint disagreement $\eta_D(\theta_i, \theta_j)$ is defined as the number of valid bits that disagree between θ_i and θ_j. The joint length $\eta(\theta_i, \theta_j)$ is the number of bits which are valid in both θ_i and θ_j. Hence,

$$\begin{aligned}
\eta_A(\theta_i, \theta_j) &= \sum_{l=1}^{D} \{m_i(l) \wedge m_j(l) \wedge \neg(x_i(l) \oplus x_j(l))\}. \\
\eta_D(\theta_i, \theta_j) &= \sum_{l=1}^{D} \{m_i(l) \wedge m_j(l) \wedge (x_i(l) \oplus x_j(l))\}. \\
\eta(\theta_i, \theta_j) &= \sum_{l=1}^{D} (m_i(l) \wedge m_j(l)).
\end{aligned} \tag{3.2}$$

The joint agreement, the joint disagreement and the joint length are related by

$$\eta_A(\theta_i, \theta_j) + \eta_D(\theta_i, \theta_j) = \eta(\theta_i, \theta_j). \tag{3.3}$$

3.2.2 Deriving a Similarity Measure

The normalized Hamming distance $\mathscr{H}(\theta_i, \theta_j)$ between two iris samples θ_i and θ_j can be expressed in terms of the joint agreement and joint disagreement as

$$\mathscr{H}(\theta_i, \theta_j) = \frac{1}{4} + \frac{1}{4} - 2\frac{\{\eta_A(\theta_i, \theta_j) - \eta_D(\theta_i, \theta_j)\}}{4\eta(\theta_i, \theta_j)}. \tag{3.4}$$

Observe that the third term in the last equation given above, $\frac{\{\eta_A(\theta_i,\theta_j)-\eta_D(\theta_i,\theta_j)\}}{\eta(\theta_i,\theta_j)}$ is the difference between the fraction of valid bits that agree and the fraction of valid bits that disagree. This provides a meaningful similarity measure between two iris codes θ_i and

θ_j. Therefore, we define the similarity measure between iris samples θ_i and θ_j as

$$\mathscr{F}(\theta_i, \theta_j) = \frac{\eta_A(\theta_i, \theta_j) - \eta_D(\theta_i, \theta_j)}{4\eta(\theta_i, \theta_j)}. \tag{3.5}$$

The scalar 4 in the denominator is just a scale factor to simplify our equations, as will become clear later.

Property: $\mathscr{H}(\theta_i, \theta_j)$ and $\mathscr{F}(\theta_i, \theta_j)$ are related by

$$\mathscr{H}(\theta_i, \theta_j) = \mathscr{F}(\theta_i, \theta_i) + \mathscr{F}(\theta_j, \theta_j) - 2\mathscr{F}(\theta_i, \theta_j). \tag{3.6}$$

3.3 Framework for Kernel Learning

In this section, we develop a framework for learning transformations of iris biometrics having desired properties. These transformations can be represented using kernel functions, and hence such techniques are called kernel learning methods [67]. The space of allowable transformations for iris biometrics and the constraints they should satisfy are described below.

3.3.1 Space of Transformations for Iris Biometrics

As discussed in Section 3.1.1, popular iris recognition techniques perform verification by matching the binary iris codes. Hence, we first need to fix the set of allowable transformations for iris codes. Boolean transformations, such as permutations, map one binary vector to another. However, learning boolean transformations satisfying desired constraints is difficult. So the class of transformations $\phi : \mathscr{B}^{2D} \to \mathbb{R}^{\mathbb{M}}$, mapping iris codes to real-valued vectors (of some dimension M) is chosen here. The corresponding kernel

function [69] is given by

$$\mathscr{K}(\theta_i, \theta_j) = \phi(\theta_i)^T \phi(\theta_j). \tag{3.7}$$

Let $\mathscr{K} \in \mathbb{R}^{N \times N}$ denote the kernel matrix, whose $(i, j)^{\text{th}}$ entry is the kernel function between θ_i and θ_j. In other words, $\mathscr{K}_{ij} = \mathscr{K}(\theta_i, \theta_j)$. Since the transformed feature vectors are real-valued, the squared Euclidean distance $\zeta_e(\cdot, \cdot)$ is used as the distance metric in the transformed space. It is related to the kernel function by

$$\begin{aligned} \zeta_e(\phi(\theta_i), \phi(\theta_j)) &= \|\phi(\theta_i) - \phi(\theta_j)\|^2 \\ &= \phi(\theta_i)^T \phi(\theta_i) + \phi(\theta_j)^T \phi(\theta_j) \\ &\quad - 2\phi(\theta_i)^T \phi(\theta_j) \\ &= \mathscr{K}_{ii} + \mathscr{K}_{jj} - 2\mathscr{K}_{ij}. \end{aligned} \tag{3.8}$$

For notational simplicity, let us denote $\zeta_e(\phi(\theta_i), \phi(\theta_j))$ by ζ_{ij}.

3.3.2 Constraints to be Satisfied

In this section, the constraints that the transformed samples must satisfy are described.

Distance preserving constraints: For the learned transformation to perform well on the test samples, the squared Euclidean distance in the transformed space should capture the distance relationships between the original iris samples. Learning transformations preserving the local distances in the original and transformed spaces is a well explored area in machine learning, called manifold learning [70, 71]. These methods are restricted to constraining the local distances, since distances between non-local points are often difficult to compute. However, since the normalized Hamming distance is a good distance measure for iris codes, we impose that the distances between all the training samples

should be preserved by the learned transformation. This can be achieved by constraining the squared Euclidean distance between the transformed vectors to be close to the normalized Hamming distance between the original vectors.

$$\zeta_{ij} \approx \mathscr{H}(\theta_i, \theta_j). \tag{3.9}$$

Application-specific constraints: Often, application -specific constraints need to be introduced into the optimization framework to obtain the desired results. For example, Weinberger *et. al.* [67] learned transformations maximizing the variance between samples. Maximum Mean Discrepancy (MMD) constraints were used for transfer learning by Pan *et. al.* [72]. Let the application specific constraints to be satisfied by the learned transformation be denoted by $\mathscr{C}(\phi) \leq 0$, where the function $\mathscr{C}(\cdot)$ depends on the constraints being imposed.

3.3.3 Kernel Learning

Having specified the space of allowable transformations and the constraints they should satisfy, the kernel learning problem can be expressed as

$$\phi^*(\cdot) = \underset{\phi:\mathscr{B}^{2D}\rightarrow\mathbb{R}^{\mathbb{M}}}{\arg\min} \sum_{\theta_i,\theta_j\in\mathscr{L}} \zeta_d(\zeta_{ij}, \mathscr{H}(\theta_i, \theta_j)) \tag{3.10}$$

subject to the constraints $\mathscr{C}(\phi) \leq 0$, where $\zeta_{ij} = \zeta_e(\phi(\theta_i), \phi(\theta_j))$, $\zeta_d(\cdot,\cdot)$ is a suitable distance measure between the squared Euclidean distance in the transformed space and the normalized Hamming distance in the original space of iris codes, and $\phi^*(\cdot)$ is the optimal mapping.

3.4 Sensor Adaptation

Having developed a general framework for learning kernel functions for iris biometrics, we now describe how it can be utilized for sensor adaptation. A sensor adaptation algorithm should reduce the sensor mismatch problem and improve the verification performance when the sensor used for enrollment differ from that used for testing. Since the algorithm has to be incorporated into existing recognition systems, it should be fast and introduce minimal changes to the existing recognition pipeline.

Let the enrollment samples be acquired using sensor S1 and testing samples using sensor S2, where S1 differs from S2 in the sensor technology or the location or type of illumination. We assume that iris samples acquired by both sensors are available for a small number of subjects. By considering the samples acquired by S2 as the target domain and those enrolled by S1 as the source domain, this becomes the standard domain adaptation problem in machine learning [73]. However, existing algorithms for domain adaptation are typically based on real-valued features. One possible solution is to convert the original iris codes from binary to real values, use an existing domain adaptation algorithm and quantize the adapted features to obtain the final iris codes for matching. However, this could lead to reduced performance due to quantization, and also lead to significant changes in the existing iris recognition systems.

Instead, we transform the binary iris codes to real-valued features using the kernel-learning framework introduced in Section 3.3. Matching is then performed using the transformed iris samples. In addition to the distance preserving constraints, the application specific constraints are incorporated for sensor adaptation, as explained below.

Inter-sensor constraints: To test samples accurately from S2 using samples enrolled by S1, the samples of S2 should be close to same-class samples in S1. Furthermore, they should be far from samples in S1 belonging to different classes. Therefore, we require that the transformation should bring samples of the same class acquired by different sensors closer, and move those from different classes farther in the transformed space [63]. These constraints are given by

$$\begin{aligned} \zeta_{ij} &\leq d_u, && \text{if } y_i = y_j \ , \ s_i \neq s_j. \\ \zeta_{ij} &\geq d_l, && \text{if } y_i \neq y_j \ , \ s_i \neq s_j. \end{aligned} \tag{3.11}$$

Intra-sensor constraints: Often sensors available for iris acquisition differ greatly in accuracy. Usually iris samples will be enrolled using an older sensor. This will have an accuracy much lower than that of the newer sensor acquiring the test samples for verification [55]. Hence, the cross-sensor performance can be limited by that of the older sensor. To handle the varying accuracies of the two sensors, additional intra-sensor constraints are introduced. For each individual sensor, they impose that the distance between same-class samples should be small, and the distance between different class samples should be large. These constraints have been used in Metric Learning [62], and will improve the performance of the older sensor. These constraints are given by

$$\begin{aligned} \zeta_{ij} &\leq d_u, && \text{if } y_i = y_j \ , \ s_i = s_j. \\ \zeta_{ij} &\geq d_l, && \text{if } y_i \neq y_j \ , \ s_i = s_j. \end{aligned} \tag{3.12}$$

Transform Learning: We can now express the transform learning problem as

$$\phi^A(\cdot) = \underset{\phi:\mathscr{B}^{2D}\rightarrow\mathbb{R}^{\mathbb{M}}}{\arg\min} \sum_{\theta_i,\theta_j\in\mathscr{L}} \zeta_d\left(\zeta_{ij}, \mathscr{H}(\theta_i,\theta_j)\right) \tag{3.13}$$

subject to the constraints

$$\zeta_{ij} \leq d_u, \qquad \text{if } y_i = y_j$$

$$\zeta_{ij} \geq d_l, \qquad \text{if } y_i \neq y_j.$$

where $\zeta_{ij} = \zeta_e(\phi(\theta_i), \phi(\theta_j))$, ζ_d is a suitable distance measure between the Euclidean distance in the transformed space and the normalized Hamming distance in the original space of iris codes, and $\phi^A(\cdot)$ is the optimal transformation for sensor adaptation.

At this point, we could have specified a parametric model for ϕ and learned its parameters by solving the optimization problem. However, it is not clear what would be a good model for ϕ, and bad choices could affect the classification performance. So, instead of a parametric approach, the optimization problem is expressed in terms of the kernel functions and the optimal kernel function is computed.

By substituting (3.8), the optimization problem can be rewritten in terms of the kernel matrix as

$$\mathscr{K}^A = \arg\min_{\mathscr{K} \in \mathscr{S}} \sum_{\theta_i, \theta_j \in \mathscr{L}} \zeta_d(\mathscr{K}_{ii} + \mathscr{K}_{jj} - 2\mathscr{K}_{ij}, \mathscr{H}(\theta_i, \theta_j)) \tag{3.14}$$

subject to the constraints, $\forall \theta_i, \theta_j \in \mathscr{L}$

$$\mathscr{K}_{ii} + \mathscr{K}_{jj} - 2\mathscr{K}_{ij} \geq d_u, \qquad \text{if } y_i = y_j$$

$$\mathscr{K}_{ii} + \mathscr{K}_{jj} - 2\mathscr{K}_{ij} \leq d_l, \qquad \text{if } y_i \neq y_j.$$

where $\mathscr{K}^A(\theta_i, \theta_j) = \phi^A(\theta_i)^T \phi^A(\theta_j)$ is the adapted kernel matrix corresponding to the optimal transformation, and $\mathscr{S}$ is the space of all positive semi definite matrices.

Direct solution: When $\zeta_d(.,.)$ is the Euclidean distance, the optimization problem above becomes convex, because it involves the minimization of a quadratic cost function subject

to linear constraints, and the global minimum can be obtained. To perform verification, the kernel function between the test samples and the training samples can be obtained by solving the problem above. Distances in the transformed space can be computed using (3.8) and used for matching, as explained in Section 3.5.3.

However, in practical applications, test iris samples are acquired at various times. Solving the optimization problem for each test sample is computationally inefficient. In the next section, we develop an efficient solution to this optimization problem based on Bregman projections [74], utilizing the similarity measure developed in Section 3.2.

3.5 Efficient solution

Substituting (3.6) in the optimization problem (3.14), the cost function to be minimized becomes

$$\sum_{\theta_i,\theta_j \in \mathscr{L}} \zeta_d(\mathscr{K}_{ii} + \mathscr{K}_{jj} - 2\mathscr{K}_{ij}, \mathscr{H}(\theta_i,\theta_j))$$

$$= \sum_{\theta_i,\theta_j \in \mathscr{L}} \zeta_d(\mathscr{K}_{ii} + \mathscr{K}_{jj} - 2\mathscr{K}_{ij}, \mathscr{F}_{ii} + \mathscr{F}_{jj} - 2\mathscr{F}_{ij}).$$

Observe that the cost function given above can be minimized by minimizing the distance between the Kernel matrix $\mathscr{K}$ and the similarity matrix $\mathscr{F}$. A suitable distance measure between the two matrices is the logDet divergence. The logDet divergence between two positive semi-definite matrices $K_1, K_2 \in \mathbb{R}^{n\times n}$ is defined as $\zeta_l(K_1,K_2) = \chi(K_1) - \chi(K_2) - \mathrm{tr}(\bigtriangledown\chi(K_2)^T(K_1-K_2))$ [74], where $\chi(K_1) = -\sum_{i,\lambda_i>0}\log\lambda_i$, λ_i is the i^{th} eigen value of K_1, $\mathrm{tr}(\cdot)$ is the matrix trace operator and $\bigtriangledown(\cdot)$ is the gradient operator. When the masks are identical, the similarity measure is a kernel function, and hence the corresponding similarity matrix $\mathscr{F}$ will be positive semi-definite. In other cases, we empirically verify

in Section 3.6.10 that the similarity matrix $\mathscr{F}$ is positive semidefinite.

The modified optimization problem is given by

$$\mathscr{K}^A = \arg \min_{\mathscr{K} \in \mathscr{S}} \zeta_l(\mathscr{K}, \mathscr{F}) \tag{3.15}$$

subject to the constraints, $\forall \theta_i, \theta_j \in \mathscr{L}$

$$\begin{aligned} \mathscr{K}_{ii} + \mathscr{K}_{jj} - 2\mathscr{K}_{ij} &\geq d_u, && \text{if } y_i = y_j \\ \mathscr{K}_{ii} + \mathscr{K}_{jj} - 2\mathscr{K}_{ij} &\leq d_l, && \text{if } y_i \neq y_j. \end{aligned}$$

where $\mathscr{S}$ is the space of all positive semi-definite matrices, $\mathscr{F}$ is the similarity matrix obtained from the training samples, $\mathscr{K}^A$ is the adapted kernel matrix and $\zeta_l(\cdot,\cdot)$ is the logDet divergence.

The optimization problem formulated above is convex as before and has a global minimum. Furthermore, the cost is a Bregman divergence [74]. An optimization problem consisting of the minimization of a Bregman divergence subject to linear inequality constraints can be solved efficiently using Bregman projections [75]. Bregman projections choose one constraint per iteration and perform a Bregman projection so that the current solution satisfies the chosen constraint. This process is repeated in a cyclic manner until convergence. Under mild conditions, it has been shown that the Bregman projection technique converges to the globally optimal solution [75]. Furthermore, as will become evident later, the optimization problem (3.15) need not be solved every time a new test sample is acquired, as is the case for (3.14).

Observe that every constraint is obtained by selecting two training samples and constraining the kernel function between them. Let $\mathscr{C} = \{(i, j)\}$ be the set of all constraints used

for sensor adaptation, where (i, j) corresponds to a constraint imposed between training samples θ_i and θ_j. Let the constraint chosen after the t^{th} iteration be formed using the t_i^{th} and the t_j^{th} data samples. Furthermore, let $e_{t_i} \in \mathbb{R}^N$ be a vector with value 1 at the t_i^{th} location and 0 otherwise. At the $(t+1)^{\text{th}}$ iteration, the Bregman update is given by [74]

$$\mathscr{K}^{t+1} = \mathscr{K}^t + \beta_{t+1} \mathscr{K}^t e_{t_i} e_{t_j}^T \mathscr{K}^t. \tag{3.16}$$

where $\mathscr{K}^0 = \mathscr{F}$, $e_{t_j}^T$ is the transpose of the vector e_{t_j}, and the scalar β_{t+1} is computed at each iteration, as explained in [74].

3.5.1 Learning Adaptation Parameters

Since only a finite number of constraints exist, and Bregman projections cyclically select each constraint for updating the kernel, the same constraint is chosen multiple times during optimization. Due to the linearity of the kernel update equation (3.16), the contribution of each constraint to the final solution can be expressed as the sum of its contribution to each iteration of the algorithm. Let τ be the total number of iterations for convergence during adaptation. Then

$$\begin{aligned} \mathscr{K}^A = \mathscr{K}^\tau &= \mathscr{K}^0 + \sum_{t=1}^{\tau} \beta_t \mathscr{K}^{t-1} e_{t_i} e_{t_j}^T \mathscr{K}^{t-1} \\ &= \mathscr{K}^0 + \sum_{(i,j)\in\mathscr{C}} \sigma_{ij} \mathscr{K}^0 e_i e_j^T \mathscr{K}^0. \end{aligned} \tag{3.17}$$

where σ_{ij}, called the adaptation parameters, represent the contribution made by the $(i, j)^{\text{th}}$ constraint to the adapted kernel. These parameters can be estimated using just the training samples during the learning stage, irrespective of testing samples.

Let $\Sigma \in \mathbb{R}^{N\times N}$ be the adaptation matrix, whose $(i, j)^{\text{th}}$ entry gives the adaptation parameter

σ_{ij}. (3.17) can be written using the matrix notation as

$$\mathscr{K}^A = \mathscr{K}^0 + \mathscr{K}^0 \Sigma \mathscr{K}^0.$$

Hence, Σ can be computed as

$$\begin{aligned}\Sigma &= (\mathscr{K}^0)^{-1}(\mathscr{K}^A - \mathscr{K}^0)(\mathscr{K}^0)^{-1} \\ &= (\mathscr{F})^{-1}(\mathscr{K}^A - \mathscr{F})(\mathscr{F})^{-1}. \end{aligned} \tag{3.18}$$

3.5.2 Sensor Adaptation during Testing

Given a testing sample θ_t, its adapted kernel function is first evaluated with all the training samples $\mathscr{K}^A(\theta_t, \theta), \theta \in \mathscr{L}$, using the adaptation parameters Σ and similarity measure $\mathscr{F}(\theta_t, \theta)$ using (3.17) as

$$\mathscr{K}^A(\theta_t, \theta) = \mathscr{F}(\theta_t, \theta) + \sum_{ij} \sigma_{ij} \mathscr{F}(\theta_t, \theta_i) \mathscr{F}(\theta_j, \theta). \tag{3.19}$$

Observe that the adapted kernel computation does not involve solving the optimization problem (3.15) for each test sample, which makes it extremely efficient.

3.5.3 Iris Matching

Given a test iris sample θ_t, its adapted kernel function values $\mathscr{K}^A(\theta_t, \theta), \theta \in \mathscr{L}$ with all the training samples are first obtained as explained above. The squared Euclidean distance in the transformed space is then computed using (3.8) as

$$\begin{aligned}\zeta_e(\phi^A(\theta_t), \phi^A(\theta)) &= \mathscr{K}^A(\theta_t, \theta_t) + \mathscr{K}^A(\theta, \theta) \\ &\quad - 2\mathscr{K}^A(\theta_t, \theta), \forall \theta \in \mathscr{L}. \end{aligned} \tag{3.20}$$

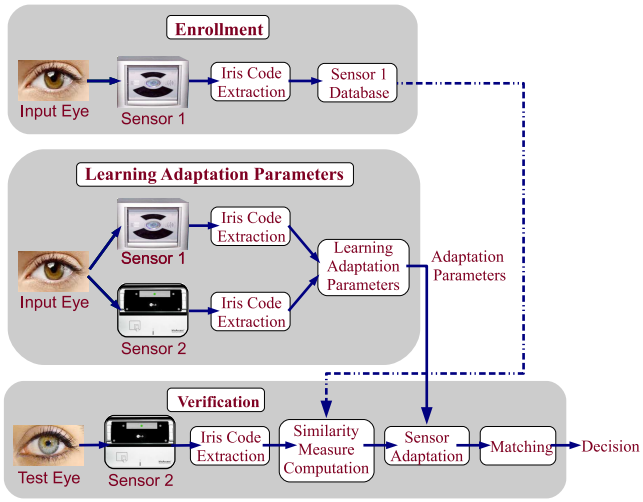


Figure 3.2: A diagram illustrating the sensor adaptation method for iris biometrics.

Verification or identification is performed as required by the application using this distance. For instance, if the squared Euclidean distance between the test sample and the sample corresponding to the claimed identity is less than a predefined threshold in the transformed space, the test sample is verified as genuine. We summarize the major steps in the proposed solution in Figure 3.2 and in Algorithms 1 and 2.

3.5.4 Extensions for Practical Systems

In this section, we describe how the proposed algorithm can handle alignment errors in iris templates, and work on non-binary features. Observe from (3.18) and (3.19) that the proposed algorithm requires only a similarity function $\mathscr{F}$, which satisfies (3.6). To apply the proposed sensor adaptation algorithm, we need to develop similarity functions satisfying relevant constraints for these scenarios.

Handling registration errors in iris templates: In-plane rotation in test iris samples is normally handled during matching by rotating one of the iris templates by different

possible values, computing the normalized Hamming distance for different rotations, and computing the effective matching distance between the two templates as the minimum of these distances. In other words, given the set of possible rotations $\mathscr{R}$ between two iris templates θ_1 and θ_2, the effective matching distance is computed as $\mathscr{H}_r(\theta_1,\theta_2) = \min_{r\in\mathscr{R}} \mathscr{H}(\theta_1, r(\theta_2))$, where the function $r(\cdot)$ represents a rotation of the iris template by a fixed angle. The corresponding similarity function can easily be derived as $\mathscr{F}_r(\theta_1,\theta_2) = \max_{r\in\mathscr{R}} \mathscr{F}(\theta_1, r(\theta_2))$. Hence, given two iris samples, we rotate one of the iris sample by possible rotations, compute the similarity measure for each rotation, and take its maximum as the effective similarity measure. Sensor adaptation is then performed using this similarity measure.

Real-valued feature representation: For real-valued features, a popular distance measure for iris recognition is the Euclidean distance. For two features $\theta_i, \theta_j \in \mathbb{R}^D$, the squared Euclidean distance is given by $\zeta_e(\theta_i,\theta_j) = \|\theta_i - \theta_j\|^2 = \theta_i^T\theta_i + \theta_j^T\theta_j - 2\theta_i^T\theta_j$. Hence, a similarity function satisfying (3.6) is the inner product function $\mathscr{F}(\theta_i,\theta_j) = \theta_i^T\theta_j$.

3.6 Experiments

In this section, we evaluate the proposed algorithm for sensor adaptation on data from two sensors, namely LG2200 and LG4000. These sensors are chosen in our experiments, since they form a real case where an older iris sensor (LG2200) was upgraded to a newer one (LG4000). The data and the implementation details are first explained. The performance of the proposed sensor adaptation algorithm is then evaluated for cross-sensor matching.

Algorithm 1: Algorithm for learning adaptation parameters.

Input: Training iris samples $\mathscr{L} = \{\theta_1, \ldots, \theta_N\}, \{y_1, \ldots, y_N\}, \{s_1, \ldots, s_N\}$

Output: Adaptation parameters $\{\sigma_{ij}, (i,j) \in \mathscr{C}\}$

1. **Similarity Measure Computation:** Compute the similarity measures $\mathscr{F}(\theta_i, \theta_j),\ \forall\ \theta_i, \theta_j \in \mathscr{L}$ using (3.5) and form the initial matrix $\mathscr{K}^0 = \mathscr{F}$.

2. **Kernel Learning:** Until convergence, update the kernel matrix using (3.16) to form the final kernel matrix $\mathscr{K}^A$.

3. **Learning Adaptation Parameters:** Using the initial similarity matrix $\mathscr{F}$ and the final matrix $\mathscr{K}^A$, compute the adaptation parameters $\{\sigma_{ij}, (i,j) \in \mathscr{C}\}$ using (3.18).

Algorithm 2: Algorithm for sensor adaptation during testing.

Input: Training iris samples $\mathscr{L} = \{\theta_1, \ldots, \theta_N\}$, adaptation parameters $\{\sigma_{ij}, (i,j) \in \mathscr{C}\}$, test sample θ_t

Output: Adapted kernel matrix $\mathscr{K}^A$

1. **Similarity Measure Computation:** Compute $\mathscr{F}(\theta_t, \theta_i),\ \forall\ \theta_i \in \mathscr{L}$ using (3.5) and form the test matrix $\mathscr{K}^0 = \mathscr{F}$.

2. **Sensor Adaptation:** Adapt the test kernel matrix using the initial test matrix and the adaptation parameters by (3.19).

Robustness of the algorithm to variations in parameters is studied. Furthermore, cross-sensor matching is performed using real-valued features. Finally, the similarity matrix $\mathscr{F}$ is empirically verified to be positive semidefinite, ensuring that the logDet divergence is a good distance measure between kernel matrix $\mathscr{K}$ and $\mathscr{F}$.

3.6.1 Iris Dataset

The iris dataset used in our experiments is the BTAS 2012 Cross-sensor Iris Competition dataset, referred to as the ND dataset, collected at the University of Notre Dame [76]. This database has iris images acquired with two sensors, namely LG2200 and LG4000. It contains about 104 Giga Bytes of iris data, collected across 27 sessions with 676 unique subjects. There are 29,939 images from the LG4000 and 117,503 original images from the LG2200. The LG2200 system has near-IR LEDs at the top, lower left, and lower right, and captures one iris at a time. The LG4000 system has near-IR LEDs on the left and right, and can image both irises of a person at the same time. The initial images taken from both sensors are of size 640 by 480 pixels. However, for the LG2200 sensor, the original images have been stretched vertically by 5% to compensate for the non-unit aspect ratio in the LG2200 acquisition system [76]. Hence, the images from the LG2200 sensor are of size 640 by 504 pixels.

3.6.2 Implementation Details

3.6.2.1 Segmentation and Feature Extraction

Iris image segmentation and feature extraction were performed using the Video-based Automated System for Iris Recogniton(VASIR) [77], an open source iris segmentation and recognition system. Evaluations on ICE 2005 and MBGC dataset have shown that VASIR can be used as a state-of-the-art baseline for still image-based iris recognition [78]. It uses contour processing and circular Hough Transform to detect the inner and outer boundaries of the iris respectively. Two ellipses are then fitted to approximate the edges of the upper and lower eyelids. The iris region is then resampled using a polar structure and mapped to a 20×240 rectangular grid. Features are then extracted by convolving it with a 1D Log-Gabor filter. The real and imaginary components of the filter response are binarized and concatenated to form a 9600 dimensional feature vector ($20 \times 240 \times 2$). Furthermore, for each of the feature dimension, a mask bit is computed, whose value is one if the corresponding rectangular grid point is inside the iris region, and zero otherwise. Hence, a 9600 dimensional mask vector is obtained to mask pixels corresponding to non iris regions like eyelids.

3.6.2.2 Evaluation Setup

Unless otherwise mentioned, for each sensor, we selected three images of both eyes from thirty subjects (180 images in total) at random to form the training data. The cross-sensor recognition performance was evaluated on the remaining subjects. Observe that this experimental setup evaluates subjects not seen during training, and hence evaluates

the generalization properties of the algorithm to unseen subjects. Furthermore, multiple images are required to be enrolled only for subjects used in training. For subjects used in the testing phase, only one image is assumed to be enrolled. To handle registration errors in iris templates, for every pair of templates, we rotate the second template two bits along the horizontal (left or right) and one bit along the vertical (upwards or downwards), as done in the VASIR system. The highest similarity value between the two templates after rotation is taken as the similarity measure, as explained in section 3.5.4.

3.6.2.3 Sensor Adaptation

During training, the adaptation parameters were computed from the training images using Algorithm 1. At first, the similarity matrix was built from all the training data using (3.5). The intra-sensor and inter-sensor constraints were then imposed, as explained in Section 3.4. The final kernel matrix was obtained using (3.16). Using the initial and final kernel matrix, the adaptation parameters were obtained using (3.18).

Parameters: Recall that the parameter d_u is the upper bound on the same class distances. Similarly d_l is the lower bound on the different class distances. In our experiments, d_u was chosen as the 20^{th} percentile of the same-class distances of the LG2200 samples. d_l was chosen as the 85^{th} percentile of the different-class distances between the LG2200 samples. The parameter γ was set as 0.1 in all our experiments. We evaluate the performance of the sensor adaptation algorithm to variations in these parameters in Section 3.6.8.

Testing: Testing was performed using Algorithm 2. For the test samples, the adapted kernel matrix was obtained using (3.19) and the squared Euclidean distance in the trans-

formed space was computed using (3.20). Verification was performed using this distance.

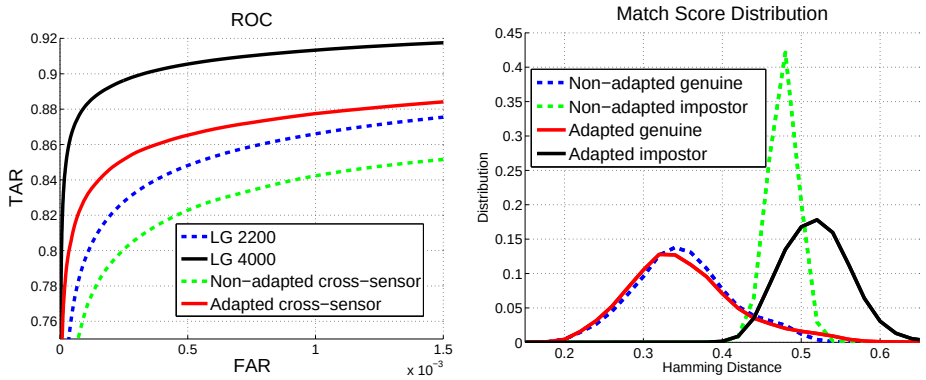
3.6.3 Cross-sensor iris recognition on the entire ND dataset.

TAR (%) at FRR=0.1%				EER (%)			
		Cross-sensor				Cross-sensor	
LG2200	LG4000	NA	**Adapted**	LG2200	LG4000	NA	**Adapted**
86.74	91.39	84.34	**87.82**	6.06	5.22	7.19	**6.09**

Table 3.1: Cross-sensor matching results for Non-Adapted (NA) case and after adaptation on the entire ND dataset.

In this section, we evaluate the proposed method on the entire ND dataset. As explained in Section 3.6.2.2, we select three images from thirty subjects in both sensors to form the training set. The training data for different sensors were chosen from the same session in this experiment. We analyze the effect of session variation on performance in Section 3.6.6. We perform pairwise matching using the learned adaptation parameters on the entire dataset. The ROC curves and the Hamming distance distributions for the non-adapted and adapted cases are shown in Figure 3.3(a) and Figure 3.3(b) respectively. Observe that the cross-sensor recognition performance is noticeably improved by adaptation, and is even better than the same sensor LG2200 results. The True Acceptance Rate (TAR) at the False Acceptance Rate (FAR) of 0.1% after adaptation is 1.08% better than the same sensor LG200 results and 3.48% better than the cross sensor performance be-

fore adaptation. Furthermore, sensor adaptation moves the genuine and impostor distance distributions apart. We attribute the performance improvement to the intra-sensor and inter-sensor constraints imposed by the proposed algorithm. The intra-sensor constraints reduce the intra-class variations between samples, and increases the inter-class variations in the transformed space, leading to better verification. The inter-sensor constraints bring the testing samples from the LG4000 sensor closer to the same class samples from the LG2200 sensor in the transformed space, improving the cross-sensor matching.



(a) (b)

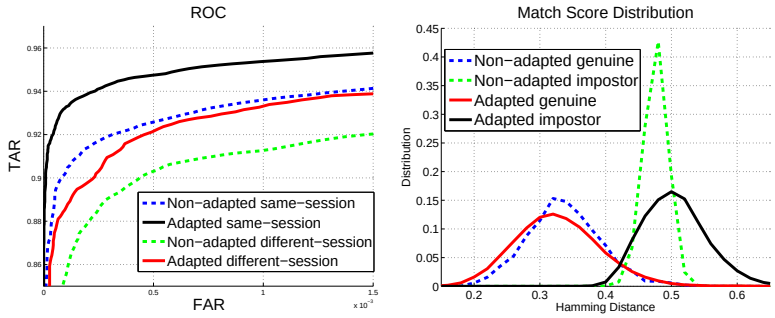
Figure 3.3: Results on the entire ND dataset. (a) The ROC curve for the adapted and non-adapted cases. (b) The Hamming distance distribution for the genuine and impostor matching before and after adaptation.

3.6.4 Cross-sensor recognition on a well segmented subset of ND dataset

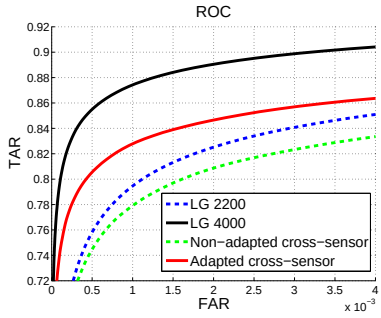
We observed that the VASIR results had segmentation errors, which reduced the same sensor recognition results of the LG2200 and LG400 sensors. In this section, we evaluate the cross-sensor recognition performance on a subset of the ND dataset, which is manually verified to be free of segmentation errors. Hence, this experiment analyzes the

behavior of the algorithm on well segmented iris data, avoiding biases due to segmentation errors. This smaller dataset consists of the left and right iris images of 123 users, thereby forming 246 unique iris signatures. For sensor LG2200, 5 images per eye were used from the same session. For sensor LG4000, two different subsets were used. The first subset was collected in the same session as the images acquired with LG2200, and consists of 246 unique irises with 3 images per eye. The second subset contained 186 unique irises and 3 images per iris. They were acquired between a month and a year after those in the LG2200 subset.

We followed the same experimental setup in Section 3.6.2.2. The ROC curves corresponding to same-session and different-session matching for the non-adapted and adapted cases are shown in Figure 3.4(a). In Table 3.2, the results are presented in the form of the Equal Error Rate (EER) and the True Verification Rate (TAR), at a False Rejection Rate (FRR) of 0.1%. Observe that the same sensor performance is better on this subset, since it does not have segmentation errors. As before, we observe that sensor adaptation improves the cross-sensor recognition performance. After adaptation, the TAR improves by 1.6% for the same-session matching, and by 1.85% for the different-session matching. For the case of matching across sessions, the cross-sensor accuracy is even better than the same sensor LG2200 accuracy. Moreover, the Hamming distance distributions in Figure 3.4(b) illustrate that adaptation moves the genuine and impostor distributions apart, leading to better discrimination between the genuine and impostor pairs. These results clearly demonstrate the performance improvement achieved by the proposed method.



(a) (b)



(c)

Figure 3.4: (a) The ROC curve for the adapted and non-adapted situations on the subset of ND dataset. (b) The Hamming distance distribution for the genuine and impostor matching before and after adaptation on the subset of ND dataset. (c) Adaptation performance using real-valued features.

3.6.5 Effect of intra-sensor and inter-sensor constraints.

In this section, we evaluate the relative importance of intra-sensor and inter-sensor constraints on the entire ND dataset. As before, we followed the evaluation setup in Section 3.6.2.2. We present the ROC curves for cross-sensor recognition in Figure 3.5 (a). Equal Error Rate (EER) and the True Verification Rate (TAR), at a False Rejection Rate (FRR) of 0.1% are provided in Table 3.3. The results demonstrate that inter-sensor constraints contribute significantly to performance improvement. This is expected, as inter-

	TAR (%) at FRR=0.1%				EER (%)			
Session	LG2200	LG4000	Cross-sensor		LG2200	LG4000	Cross-sensor	
			NA	**Adapted**			NA	**Adapted**
Same	97.78	99.52	94.13	**95.73**	1.46	0.36	2.85	**2.26**
Diff.	93.53	97.61	91.93	**93.78**	3.04	1.32	3.56	**2.87**

Table 3.2: Cross-sensor matching results on the subset of ND dataset for the Non-Adapted (NA) and Adapted cases.

sensor constraints are responsible for reducing the variations between the sensors. Furthermore, combining the inter-sensor and intra-sensor constraints gave the highest accuracy.

3.6.6 Effect of session variations.

If the training data from different sensors are collected in different sessions, it is possible that the proposed method will learn the session variations too, along with the sensor variations. To evaluate the effect of these session variations, we used training data for each sensor from a different session. We then evaluated the cross-sensor performance on new sessions unseen during training. All other evaluation settings were identical to that in Section 3.6.3. We present the ROC curves in Figure 3.5 (b) and the corresponding results in Table 3.4. As expected, when the training data for each sensor is chosen from a differ-

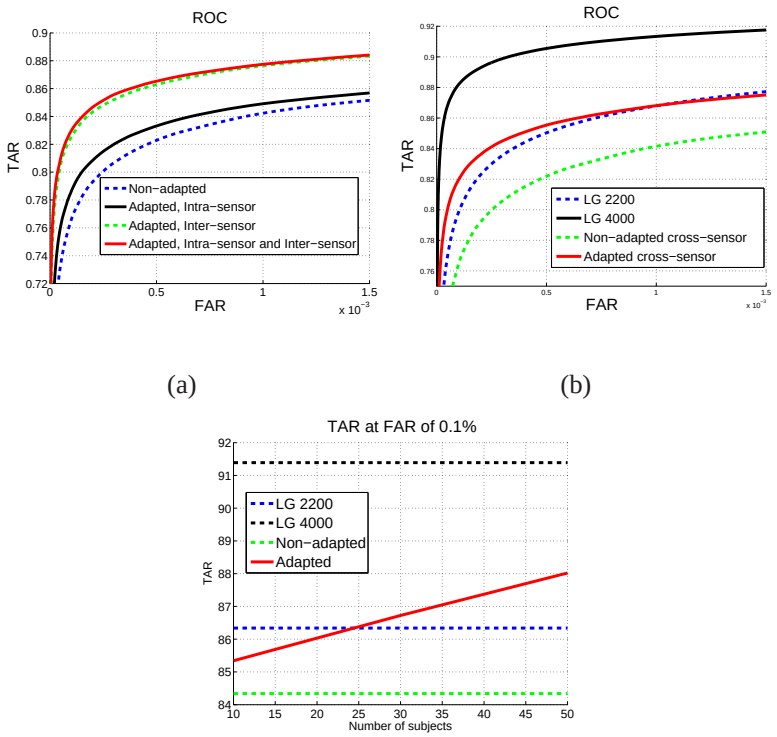


Figure 3.5: (a) Results of intra-sensor and inter-sensor constraints. (b) Effect of session variations on cross sensor recognition. (c) Effect of training size on cross sensor recognition.

ent session, the true acceptance rate after adaptation is reduced from 87.82% to 86.87%. However, this accuracy is still better than that of the LG2000 same sensor recognition accuracy of 86.81% and the non adapted cross sensor accuracy of 84.27%. This experiment demonstrates that the proposed method generalizes across unseen sessions.

3.6.7 Number of subjects during training.

In this section, we analyze the effect of the size of training data on cross-sensor recognition accuracy. We plot the True Acceptance Rate (TAR) at a False Acceptance Rate(FAR)

TAR (%) at FRR=0.1%				EER (%)			
	Adapted				Adapted		
NA	Intra-sensor	Inter-sensor	**Both**	NA	Intra-sensor	Inter-sensor	**Both**
84.34	84.90	87.73	**87.82**	7.19	7.19	6.14	**6.09**

Table 3.3: Effect of intra-sensor and inter-sensor constraints on cross-sensor recognition for the Non-Adapted (NA) and adapted cases with intra-sensor, inter-sensor and their combination.

of 0.1% with varying number of subjects for training in Figure 3.5 (c). All other evaluation settings are identical to those explained in Section 3.6.2.2. Observe that even with ten subjects, the cross-sensor recognition accuracy after adaptation is better than that of the non-adapted case. Furthermore, the cross-sensor recognition accuracy improves with more training data. This is expected as more constraints are available for learning as training data increases.

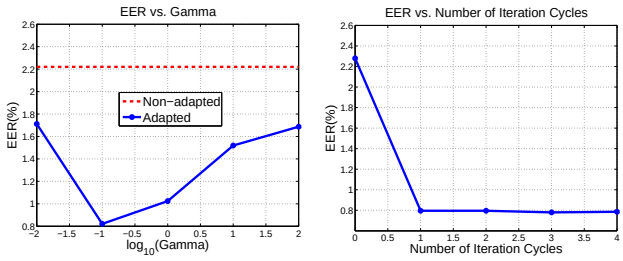
3.6.8 Robustness to Parameters

The parameters of the proposed algorithm are the parameter γ, the number of iterations τ of the Bregman update, and the distance threshold d_u and d_l. We analyze the robustness of the sensor adaptation algorithm to variations in these parameters in this section. In Figure 3.6(a), the EER corresponding to different values of the parameter γ is shown. While the best performance is obtained using $\gamma = 0.1$, the proposed algorithm improves

TAR (%) at FRR=0.1%				EER (%)			
		Cross-sensor				Cross-sensor	
LG2200	LG4000	NA	**Adapted**	LG2200	LG4000	NA	**Adapted**
86.81	91.39	84.27	**86.87**	6.04	5.22	7.55	**6.89**

Table 3.4: Cross-sensor matching results on unseen sessions for the Non-Adapted(NA) and adapted cases.

the equal error rate for a wide range of γ, illustrating its robustness to the parameter. We refer to performing Bregman projections over all the constraints once as an "iteration cycle". Figure 3.6(b) shows the variation in EER for different number of iteration cycles in the training stage. It indicates that the proposed algorithm converges quickly after all the constraints have been visited once, and further update does not change the performance. Furthermore, we observed little variation in cross-sensor matching performance with significant variations in the distance thresholds d_u and d_l.



(a) (b)

Figure 3.6: Variation of verification accuracy during testing with (a) parameter γ and (b) number of iteration cycles in the learning algorithm.

3.6.9 Incorporating Real-valued Features

In this section, we evaluate the cross-sensor recognition performance using real-valued features on the entire ND dataset. Phase of the 1D Log-Gabor features was used as the real-valued feature in our experiment, since it can be obtained directly from the VASIR system. However, the proposed algorithm can be applied to other features also, as explained in section 3.5.4. After performing iris segmentation and unwrapping, the 1D Log-Gabor filter was applied and the phase of the filter output at each pixel was computed. The kernel computation was performed using the linear kernel function, as explained in Section 3.5.4. Squared Euclidean distance between the transformed features was used for matching. The ROC curves are presented in Figure 3.4 (c), and a summary of the results appear in Table 3.5. As in the case of binary features, sensor adaptation improves the cross-sensor matching accuracy significantly. Also, the true acceptance rate after adaptation is better than the LG2200 same sensor recognition performance.

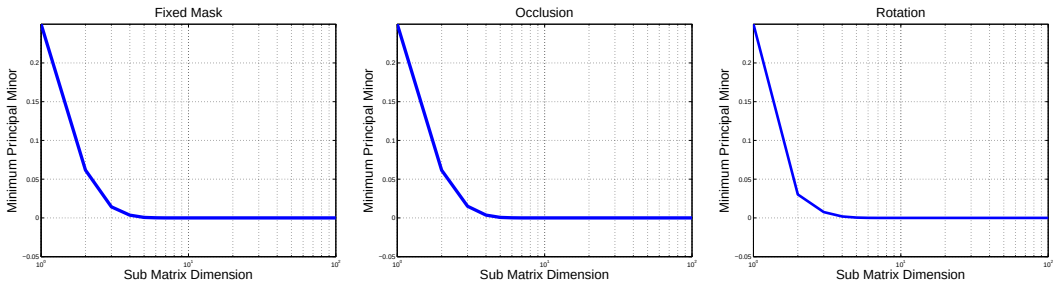
TAR (%) at FRR=0.1%				EER (%)			
		Cross-sensor				Cross-sensor	
LG2200	LG4000	NA	**Adapted**	LG2200	LG4000	NA	**Adapted**
79.59	87.50	78.14	**82.89**	7.4	5.55	8.55	**7.57**

Table 3.5: Cross-sensor matching results using real-valued features on the entire ND dataset for the Non-Adapted(NA) and Adapted cases.

3.6.10 Empirical Verification of Positive Semidefiniteness of the Similarity Measure

To use the logDet divergence in Section 3.5, the similarity matrix $\mathscr{F}$ should be positive semidefinite. To verify this empirically, one could check whether the eigen values of the similarity matrix are non-negative. However, eigen value computation of large matrices is often imprecise due to numerical errors. Hence, we adopt the Principal Minor Test [79]. By definition, the k^{th} principal minors of a matrix are the determinants of the submatrices formed by deleting any $n-k$ rows and the corresponding columns of that matrix. By the Principal Minor Test, a necessary and sufficient condition for a matrix to be positive semidefinite is that all possible principal minors of the matrix are non-negative.

Using 1,622 iris samples acquired in both LG2200 and LG4000 sensors, as explained in Section 3.6.4, we construct the similarity matrix corresponding to the fixed mask, varying mask due to occlusion and the rotation cases. For a given matrix with n rows and a particular submatrix dimension k, there are $\binom{n}{k}$ principal minors, which increases exponentially with k. Given the large number of possible minors, for each submatrix dimension, we randomly choose a fixed number of principal minors (chosen as 100 in our experiments) and compute their determinant. We plot the minimum of the randomly chosen minors in Figure 3.7. While there are 1,622 submatrix dimensions, we show the initial 100 dimensions in Figure 3.7 for clarity. The minimum of the chosen minors were non-negative for each submatrix dimension, indicating that all the chosen minors are non-negative. This empirically verifies that the similarity matrix $\mathscr{F}$ is positive semidefinite.



(a) (b) (c)

Figure 3.7: Plot of the minimum principal minor for each submatrix dimension of the similarity matrix, for the (a) fixed mask, (b) occlusion and (c) rotation cases. Observe that the minimum principal minors are non-negative for all submatrix dimensions, empirically verifying that the similarity matrix is positive semidefinite.

3.6.11 Hardware and Computational Complexity

In the proposed sensor adaptation method, we have to find the squared Euclidean distance between each test sample and the enrolled samples in the transformed space during testing. The additional steps introduced are the computation of the initial kernel using (3.6), the adaptation of the kernel using (3.19), and the calculation of the adapted Hamming distance using (3.20). However, observe that these three steps are simple linear operations, and introduce limited overhead to the original iris recognition system. Furthermore, the only additional components required in the system are adders and multipliers, and can easily be incorporated into existing systems.

	Non-adapted	**Adapted**
Time(s)	25.5	**27.9**

Table 3.6: Comparison of the testing time for the non-adapted and adapted cases.

Asymptotic Analysis: We analyze the computational complexity of the proposed method during testing below. Let D be the dimension of iris samples, and N_{Tr} be the number of training samples. In one-versus-one verification case, the test sample has to be matched with a single enrolled sample. In this case, non-adapted method requires the computation of a single normalized Hamming distance, which is an $\mathscr{O}(D)$ operation. For the adapted algorithm, similarity measure has to be computed between the testing sample and all the training samples. Furthermore, these values have to be combined with the adaptation parameters using Equation (3.19). So the total computational complexity of the proposed method is $\mathscr{O}(D * N_{Tr} + N_{Tr})$. Since template dimension and training samples are fixed, both the methods run in constant time asymptotically.

Now let us consider the case, when the test sample has to be matched with N_E samples in the gallery. Normally $N_E \gg N_{Tr}$. The computational complexity of the non-adapted algorithm is $\mathscr{O}(DN_E)$, since it has to compute the normalized Hamming distance of the test sample with N_E enrolled samples. For the proposed method, similarity measures between the enrolled samples and the training samples can be precomputed, along with the adaptation parameters. So the computational complexity during testing is due to the computation of similarity measures between the test samples and the N_{Tr} training samples, and their combination with the adaptation parameters. So the total complexity during testing for sensor adaptation is $\mathscr{O}(DN_{Tr} + N_E N_{Tr}) = \mathscr{O}(N_E N_{Tr})$. Hence, when a query image has to be compared with multiple enrolled samples, the computational complexity of both the non-adapted and adapted methods vary linearly with the number of enrolled samples.

Empirical Evaluation: In this section, we compare the testing time for the non-adapted

and adapted cases on an Intel Dual Core 2.33GHz processor. In the non-adapted case, for each LG4000 sample, we record the time for iris image segmentation, feature extraction, and matching with all the samples in the LG2200 dataset. In the adapted case, along with the segmentation and feature extraction times, the time for computing the squared Euclidean distance in the transformed domain and matching with the LG2200 samples are included. The experiment is run 10 times and the average testing times are reported in Table 3.6. As can be observed, the sensor adaptation algorithm leads to only a small increase in the execution time.

Chapter 4

Temporal Inference from Human Pose

4.1 Introduction

Automatic analysis of visual data involving humans is an important area in computer vision [80], which is useful in entertainment, human computer interaction and security. Since traditional applications like people tracking [81] and activity recognition [82] are video-based, motion cues plays an important part in these applications. However, with the availability of personal photo collections and sports images, analysis of humans in still images is gaining importance recently. These image-based applications do not have explicit motion cues, and are currently limited to using just the appearance cues [83, 84]. This leads us to an interesting question: Can implicit motion cues be extracted from still images of humans, and used to aid visual analysis?

Estimating motion without multiple images seems impossible at first. However, extensive studies in psychology have shown that information about posture of the human body plays a vital role in biological motion perception [85, 86]. Experiments of Hirai *et al.* [87] demonstrated that destroying the body structure led to a higher reduction in motion perception in humans when compared to destroying the temporal structure of motion. Furthermore, humans can easily anticipate the future motion of actors from their current body configuration [88]. As an example, consider predicting the future motion of humans from their current poses in Figure 4.1. In the first case, one can easily infer the future motion of

the human, namely, the right hand moving forward with the left hand moving backward, and the legs moving in the opposite direction. In the second case, we expect the hands to move and other parts to remain still. However, predicting the exact motion of the human is not easy. Two possible future motions corresponding to the same current human pose is shown in the right in case 2. Thus, the first pose conveys more information about the future trajectory of motion compared to the second. In this work, we refer to this information conveyed by humans poses about their future motion as the "dynamic information" in the pose. Furthermore, estimation of motion information from still images of humans is termed as "dynamic inference".

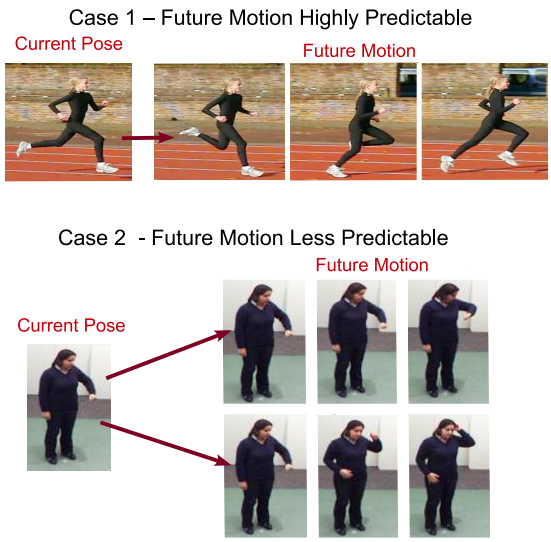


Figure 4.1: Consider predicting the future motion of the human from the current poses given in the left, for each case above. In case 1, the future motion can be easily predicted. However, the exact future motion is not obvious in case 2. Possible future motions are shown in the right.

Dynamic information in human poses can aid computer vision systems in multiple ways.

Similar to biological systems, vision systems can utilize this information to efficiently predict the future motion of human users. Human motion prediction is also useful in robotics. For instance, in robotic applications like "assistance to manipulation", robots often assist humans or manipulate the same object as humans. In such applications, accurate prediction of human motion can improve robotic performance, as empirically verified by Jarrasse *et al.* [89]. Dynamic information in poses can also improve activity recognition from still images and aid the synthesis of realistic human motion. The latter is useful in applications involving humanoid robots and animation. Additionally, poses with high dynamic information also reveal the "story" in a large number of adjacent frames, making them ideal candidates for key-frames in video summarization applications. This idea of using human poses to convey the "story" has been exploited by artists in paintings and sculptures. Japanese Manga images in Figure 4.2 is a case in point, where Hokusai pioneered the technique of conveying motion using physically unstable human configurations.

Motivated by the above, we develop a computational model to infer the "next move" from still images of humans. Our goal is to predict the future motion of a human given a single pose and quantify the extent to which it is constrained by a given pose. We emphasize that the input to our algorithm is just a single human image and the final goal is to predict the motion of the human and not the type of action performed.

Contributions: We make the following contributions in this work. We explore the potential of the implicit dynamic information conveyed by a human pose. We develop a probabilistic framework to model it. Using this framework, we estimate the amount of dynamic information conveyed by a pose and predict the probable future motion. The

Figure 4.2: Database of 45 Hokusai Manga Images. The functional Magnetic Resonance Imaging (fMRI) studies by Osaka *et al.* [2] illustrated that the dancer images on the left in unstable poses activated the motion sensitive visual cortex in humans, indicating that humans can perceive the implied motion in these images. However, the priest images on the right in stable poses elicited low responses of implied motion in humans. We use this dataset to validate the proposed computational model. in our experiments.

proposed method requires limited manual supervision since it uses unlabeled and unsegmented human videos for training, and can easily be implemented. We demonstrate the usefulness of the estimated dynamic information in a variety of vision applications like human motion prediction, activity recognition and video summarization.

Organization of the chapter: A brief review of related works is presented in Section 4.2. The proposed framework for extracting dynamic information in human pose is introduced in Section 4.3. Computer vision applications which benefit from the extracted dynamic information are enumerated in Section 4.4. We empirically evaluate the proposed technique in Section 4.5.

4.2 Related Work

Visual analysis of humans from images and video is one of the central problems in computer vision [80]. Some of the tasks associated with it are human detection, recognition, tracking, articulated pose estimation and activity recognition. In applications involving videos like human tracking, motion cue plays an important part [81]. However, still image-based applications like analysis of commercial photographs, sports images or newspaper images lack such motion cues, making them more challenging. In this work, we focus on such applications, where no explicit motion cue is available.

For still images, two problems which have received a great deal of attention in recent years are human detection and articulated pose estimation. Below, we briefly describe some of the popular methods. Numerous works have looked at finding pictures of humans [90], localizing people in still images [91], and pedestrian detection [92]. Dalal and Triggs proposed the Histogram of Oriented Gradients (HOG) [5], a popular gradient based feature for human detection. Zhu *et al.* [93] advanced HOG descriptors by combining HOG and AdaBoost to select the most suitable block for detection. Tuzel *et al.* [94] developed the covariance descriptor for human detection. Felzenswab *et al.* [95] developed discriminatively trained part-based models for human detection, using latent SVMs. For 2D pose estimation of humans, Ramanan *et al.* [96] presents an iterative parsing process for pose estimation of articulated objects. Andriluka *et al.* [7] developed a general framework-based on pictorial structures for human detection and 2D pose estimation. Bourdev and Malik [97] developed poselets capturing the 2D appearance and 3D joint position of humans, which has been utilized for human detection, segmentation and

pose estimation.

Recently, researchers have recognized that still images of humans contain not only information about the configuration of body parts, but also higher level information like the action being performed. This has led to the development of action recognition algorithms from still images. Thurau *et al.* [83] recognized human actions from still images and video, by representing actions as a histogram of pose primitives, and using histogram matching for recognition. Ikizler *et al.* [98] represented the human pose using histogram of rectangular regions and used SVMs for classification. [99] used oriented rectangular patches extracted from the human silhouette to represent the action and histogram matching for recognition. Human pose in the query image was considered as a latent variable in [100]. Latent SVM was used for recognizing activities in this work. However, these techniques are often applicable only for simple actions, since complex activities cannot always be captured by a single pose. Nevertheless, even poses belonging to complex activities often provide information about the local motion trajectory. For instance, consider the pose π_2 in Figure 4.3. While it is easy to infer that the person is bending down, it is difficult to predict the subsequent activity (for example sitting down or picking up a ball). In this work, we focus on estimating this motion information associated with the human pose in still images.

Another line of research which motivated our work is motion estimation from still images of natural scenes. Roth and Black [101] learned the prior probability of motion fields from still images of natural scenes using a Markov random field model. Their experiments demonstrated that the learned motion prior capture the rich spatial structure found in natural scenes, and can also improve motion estimation accuracy in test videos. Liu *et.*

al [102] proposed SIFT flow, a method to densely align scene images by matching densely sampled pixel-wise SIFT features, while preserving continuity. Motion of pixels in query images were then predicted by transferring SIFT flow from similar training images. Yuen and Torralba [103] learned the probability density of local motion trajectories in a non-parametric manner at each pixel location, and used samples from the density to estimate the motion trajectories in query scene images. These methods capture only the local structure of the scene, and not the influence of the global scene on the ensuing motion. Hence, they are not directly applicable to human motion prediction, where future motion is dependent on the global pose of the human. On the other hand, we directly model the relationship between the human pose and the future motion of the human body in this work.

4.3 Dynamic inference from a human pose

Before developing a computational model, we first analyze the physical evidence for the existence of dynamic information in this section. Starting at a particular pose, the future motion of the human body is constrained by numerous factors. The mechanics of body joints prevent arbitrary motion of the body. Laws of physics like gravity and momentum also limit the future moves of the human. Furthermore, every realistic pose is part of a human activity with a well defined objective. These constraints on the future motion of the human body are responsible for the dynamic information associated with a particular pose. Furthermore, the set of possible future motions vary widely between different human poses, as can be observed from Figure 4.1. Here, in case 1, the future motion of the

human is highly constrained. However, in case 2, numerous future motions are possible starting at the same pose. Hence, poses differ in the amount of dynamic information they possess.

To model the relationship between human pose and the ensuing motion, we have to decide the representation for the pose, the future motion and the space of allowable models. For pose representation, a popular choice is the articulated model [7], which represents the human body as collection of parts and learns the appearance model for each part. This model however has an explicit training stage and is not robust to unseen poses. Hence, we choose the simpler HOG-based model [83], representing the human pose using the HOG features extracted from the bounding box. This avoids the need for training models for pose estimation, can generalize to new poses and is robust to errors in the estimated pose parameters.

Given a human pose, there is a set of possible motion trajectories originating from it, and the exact future motion is uncertain. This is evident in case 2 in Figure 4.1, where two possible future motions starting from the same pose are shown in the right. To capture this uncertainty, we develop a probabilistic framework, estimating the conditional probability distribution of subsequent human motion given a pose. Once this distribution is obtained, one can compute different statistics, which ultimately yields quantities of interest. For instance, two useful statistics are the mode of the distribution and its entropy. Given a single pose, the mode of the conditional distribution gives us the most probable temporal evolution of poses. The entropy of this distribution measures the uncertainty in these future sequences. The work of Kerzel [104] shows that this uncertainty (unpredictability) provides a measure of the amount of dynamic information perceived by humans in a pose.

The higher the predictability of motion from a pose, the higher the dynamic information it conveys.

To develop a probabilistic model, we first need to define the space of predictions. Firstly, from a stable pose such as case 2 in Figure 4.1, the set of possible human motions that can follow is extremely large. Further, even for predictable poses where the set of future motions is potentially constrained, there is an equivalence class of future motions differing only in the rate of execution. Hence, we need a representation of motion invariant to the rate of execution. Considering the above, we first model human activities as a sequence of movements called action segments, separated by "ballistic" boundaries [105]. These movements are natural units of human actions, typically comprising an initial acceleration of limbs towards a target followed by deceleration to stop the movement. Figure 4.3 shows a simple illustration of the ballistic boundaries. Here, the ballistic boundaries highlighted in red separate the "picking up" action into two action segments, namely the "bending down" action segment and the "getting up" action segment. Vitaladevuni *et al.* [105] have been developed computational models to automatically extract ballistic motion boundaries from videos. By viewing actions as separated by ballistic motion boundaries, we can restrict the scope of the motion prediction problem to predicting statistics over future action segments, which are shorter in duration. In addition, since ballistic boundaries are robust to the rate of execution, the estimated statistics become invariant to the rate of execution of the action.

Before developing the model, we first introduce the notation and elements of our framework. Let π_i represent the i^{th} pose and $\Pi = \{\pi_i, i = 1, \ldots M\}$ be the set of all possible human poses. Similarly, let ϕ_i represent the i^{th} action segment and $\Phi = \{\phi_i, i = 1 \ldots N\}$

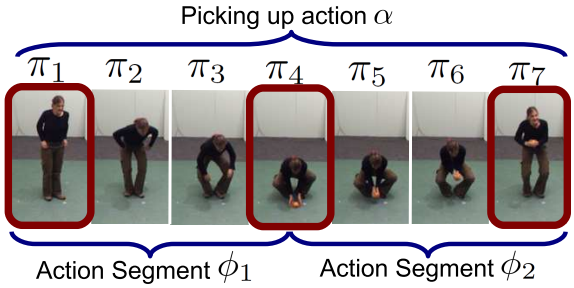


Figure 4.3: Illustration of ballistic boundaries for the "picking up" action. The three ballistic boundaries π_1, π_4 and π_7, highlighted in red, divide the action α into two action segments ϕ_1 and ϕ_2.

be the set of all possible action segments. Any action α is a temporally ordered sequence of action segments $\left[\phi_{\alpha_1}, \ldots \phi_{\alpha_{t(\alpha)}}\right]$, where each action segment ϕ_k is itself a temporally ordered sequence of individual poses $\left[\pi_{k_1}, \ldots \pi_{k_{t(k)}}\right]$. We illustrate these notations for the simple action of picking up a ball in Figure 4.3. This action α consists of two action segments $[\phi_1, \phi_2]$. Action segment ϕ_1 is a temporally ordered sequence of poses $[\pi_1, \pi_2, \pi_3, \pi_4]$. Similarly, action segment ϕ_2 is a temporally ordered sequence of poses $[\pi_4, \pi_5, \pi_6]$.

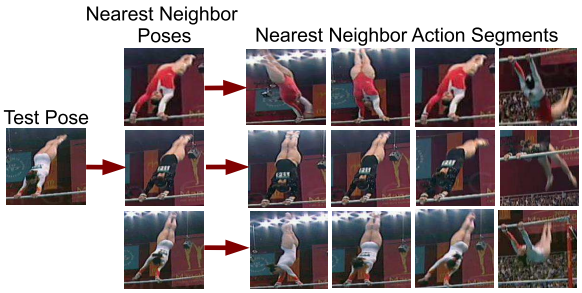


Figure 4.4: Nearest neighbor poses and the associated action segments corresponding to a test pose.

Let $\mathscr{P}(\phi|\pi)$ denote the conditional probability that a given pose π occurred in an action

segment ϕ . As discussed earlier, the uncertainty in the temporal evolution of poses starting from π is low, if it has high dynamic information. In an information theoretic framework, this uncertainty can be measured by the entropy $\mathscr{H}(\phi|\pi)$ of the conditional distribution of an action segment given a pose.

$$\mathscr{H}(\phi|\pi) = -\int_{\phi\in\Phi} \mathscr{P}(\phi|\pi)\log(\mathscr{P}(\phi|\pi))d\phi \tag{4.1}$$

This motivates our measure, Degree of Dynamic Information (DDI) of a pose, which can be computed as

$$\text{DDI}(\pi) = \exp[-\mathscr{H}(\phi|\pi)] \tag{4.2}$$

where the negative exponent captures the inverse relationship between uncertainty in the temporal evolution of poses starting from π and the amount of dynamic information in π. Another piece of valuable information that can be immediately obtained from $\mathscr{P}(\phi|\pi)$ is the most probable action segment $\hat{\phi}$ that contains the pose π.

$$\hat{\phi}(\pi) = \underset{\phi\in\Phi}{\arg\max}\ \mathscr{P}(\phi|\pi) \tag{4.3}$$

Similarly, given a start pose π_s and an end pose π_e, we can obtain the most probable pose trajectory as

$$\hat{\phi}(\pi_s,\pi_e) = \underset{\substack{\phi\in\Phi \\ \phi=[\pi_s,\ldots,\pi_e]}}{\arg\max}\ \mathscr{P}(\phi|\pi_s,\pi_e) \tag{4.4}$$

Having defined the two terms using $\mathscr{P}(\phi|\pi)$, the question now turns to the estimation of this density. Explicitly modeling this density and estimating its parameters from finite training data is extremely difficult and prone to overfitting due to the large variations in

humans poses and future motions in unconstrained settings. Hence, we adopt the data-driven approach, which has become very popular in recent years [102, 106, 107]. This approach advocates transferring information from a rich training database to the specific query under consideration, instead of learning a general function applicable to all queries. Such methods have shown significant promise in solving otherwise difficult tasks such as scene alignment [102], geo-localization [106], scene completion [108], scene parsing [107] and object matching [109].

Given a test post π_s, we estimate $\mathscr{P}(\phi|\pi_s)$ directly from the training data. This estimate is then used to compute the amount of associated dynamic information $\mathrm{DDI}(\pi_s)$ and the most probable action segment $\hat{\phi}(\pi_s)$. We explain this approach in detail below.

4.3.1 Estimation of Conditional Distribution

Instead of developing a functional form for $\mathscr{P}(\phi|\pi_s)$, we compute this probability whenever we encounter a test pose π_s. Our training data consist of videos of human actions. Let $\mathscr{D}$ denote the database of all the poses, which are extracted from these videos. By applying the temporal segmentation algorithm of Vitaladevuni *et al.* [105], these videos are divided into action segments separated by ballistic boundaries. Given a test pose π_s, we find all the instances of the pose in the database $\mathscr{D}$ and denote this set by $\mathscr{N}_{\pi_s}$. In our experiments, nearest neighbors of the test pose π_s in the database $\mathscr{D}$ are used to form the set $\mathscr{N}_{\pi_s}$. Note that every pose $\pi \in \mathscr{D}$ is a part of an action segment $\phi \in \Phi$. This implies that every pose $\pi_r \in \mathscr{N}_{\pi_s}$ has an associated action segment $\phi(\pi_r)$. Let $\mathscr{N}_{\phi(\pi_s)}$ be the set of action segments corresponding to the poses in $\mathscr{N}_{\pi_s}$. We illustrate the nearest neighbor

poses and the associated action segments for a test pose in Figure 4.4.

$\mathscr{N}_{\phi(\pi_s)} = \{\phi(\pi), \pi \in \mathscr{N}_{\pi_s}\}$ can be considered as samples from the density $\mathscr{P}(\phi|\pi_s)$. Hence, sample-based density estimation techniques can be adopted to estimate $\mathscr{P}(\phi|\pi_s)$ given $\mathscr{N}_{\phi(\pi_s)}$. However, such techniques cannot be applied directly on the space of action segments Φ due to two reasons. First of all, action segments can differ in the number of frames. Hence, a direct representation in terms of the associated pixels lead to vectors of different dimensionality. Secondly, this direct representation in terms of the associated pixels is high dimensional. Learning models from higher dimensional data is often impractical, and has lead to the development of alternate low dimensional representations for the data [110]. Hence we adopt a parametric approach, where the action segments are compactly represented by a low dimensional dynamical model.

Modeling Action Segments: In this work, we employ the Linear Dynamical System (LDS) [111], a popular dynamical model in computer vision. This model has been successfully used to represent actions, dynamic textures and human joint angle trajectories. However, it is important to note that the proposed framework of dynamic inference is a general one, and can be applied to other models also. For an action segment ϕ, the LDS model is described by

$$z_\phi(t+1) = A(\phi)z_\phi(t) + v_\phi(t), v_\phi(t) \sim N(0, \Xi) \tag{4.5}$$

$$y_\phi(t) = C(\phi)z_\phi(t) + w_\phi(t), w_\phi(t) \sim N(0, \Theta)$$

where $z_\phi(t) \in \mathbb{R}^p$ is the hidden state vector for the t^{th} frame in the action segment ϕ, $y_\phi(t) \in \mathbb{R}^d$ are the features extracted from t^{th} frame, $A(\phi) \in \mathbb{R}^{p \times p}$ is the transition matrix, $C(\phi) \in \mathbb{R}^{d \times p}$ is the measurement matrix. $v_\phi(t)$ and $w_\phi(t)$ are the noise components,

which are modeled as Gaussian with mean zero and covariances Ξ and Θ respectively. $A(\phi)$ is constrained to have eigen vectors inside the unit circle, while $C(\phi)$ is constrained to be orthonormal. Hence, the parameters of the LDS model, namely $(A(\phi), C(\phi))$ do not lie on the Euclidean space. For comparison of actions, a commonly used distance metric is the subspace angles between the column spaces of the corresponding observability matrices. The 'observability' matrix of an action segment ϕ is given by

$$\Omega^\top(\phi) = \left[C(\phi)^\top, (C(\phi)A(\phi))^\top, \ldots, (C(\phi)A(\phi)^{m-1})^\top, \ldots\right]$$

It is an infinite dimensional matrix, which can be approximated by the finite matrix

$$\hat{\Omega}^\top(\phi) = \left[C(\phi)^\top, (C(\phi)A(\phi))^\top, \ldots, (C(\phi)A(\phi)^{m-1})^\top\right]$$

Note that $\hat{\Omega}^\top(\phi) \in \mathbb{R}^{n\times d}$, where $n = mp$. Hence the column space of $\hat{\Omega}^\top(\phi)$ is a d-dimensional subspace in $\mathbb{R}^n$, which constitute the Grassmann manifold $\mathscr{G}_{n,d}$. For notational simplicity, we denote the observability matrices $\hat{\Omega}(\phi)$, $\hat{\Omega}(\phi_i)$ and $\hat{\Omega}(\phi_j)$ by Ω, Ω_i and Ω_j respectively. Then a natural distance metric between these action segments ϕ_i and ϕ_j is given by [112].

$$\zeta^2(\Omega_i, \Omega_j) = d - tr(\Omega_j^T \Omega_i \Omega_i^T \Omega_j) \tag{4.6}$$

Density Estimation on the Grassmann Manifold: Using $\mathscr{N}_{\phi(\pi_s)}$, the set of samples from $\mathscr{P}(\phi|\pi_s)$, we now estimate the conditional density using non parametric density estimation techniques [113], as

$$\hat{\mathscr{P}}(\phi|\pi_s) = c_1 \sum_{\phi_i \in \mathscr{N}_{\phi(\pi_s)}} \Psi(M^{-\frac{1}{2}}(I_d - \Omega_i^\top \Omega\Omega^\top \Omega_i)M^{-\frac{1}{2}}) \tag{4.7}$$

where $\Psi(T) = \exp(\mathrm{tr}(-T))$ for $T \in \mathbb{R}^{d\times d}$, tr(.) is the matrix trace operator, $M \in \mathbb{R}^{d\times d}$ is a smoothing matrix and c_1 is a normalization factor.

4.3.2 Statistical Inference on the Estimated Density

Having formulated the conditional density for the action segment given the test pose, we now estimate statistics of interest from it. The block diagram of the proposed method in shown in Figure 4.5, and the details are explained in Algorithm 3.

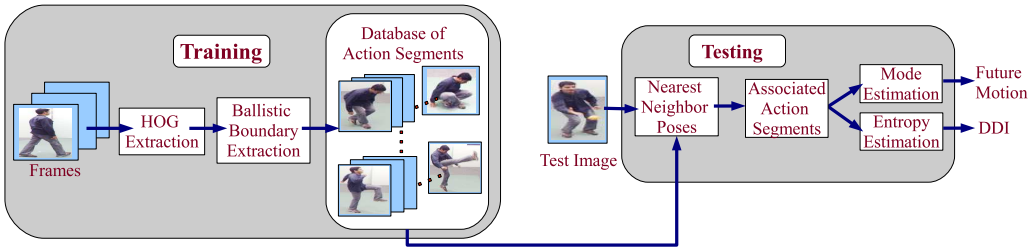


Figure 4.5: Block diagram demonstrating the various steps in the proposed method.

Mode Estimation: Given a pose π_s, the likely future motion can be predicted by finding the most probable action segment $\phi^*(\pi_s)$, which is the mode of the distribution $\mathscr{P}(\phi|\pi_s)$. Non-parametric techniques have been recently developed for mode seeking on analytic manifolds [114, 115]. In particular, Cetingul and Vidal [115] computes the mode on the Grassmann manifold using iterative optimization. It intrinsically locates the modes of the distribution via consecutive evaluations of a mapping. For Grassmann manifold, these evaluations constitute an efficient gradient ascent scheme, which avoids the computation of expensive exponential mappings. However, this algorithm will only compute the LDS parameters of the most probable action segment. It is not possible to generate the frames of the action segment from the LDS parameters. Hence, in applications where a valid action segment with high probability of occurrence is required, a more efficient scheme

is to directly select the action segment with the highest conditional density from $\mathcal{N}_{\phi(\pi_s)}$.

$$\hat{\phi}(\pi_s) = \underset{\phi_i \in \mathcal{N}_{\phi(\pi_s)}}{\arg\max}\ \hat{\mathcal{P}}(\phi_i|\pi_s) \tag{4.8}$$

By similar analysis, we can also obtain $\hat{\phi}(\pi_s, \pi_e)$, the most probable pose trajectory given a start pose π_s and an end pose π_e, by using the samples from $\mathcal{N}_{\phi(\pi_s,\pi_e)}$. Here, $\mathcal{N}_{\phi(\pi_s,\pi_e)}$ denote the set of training action segments, whose start and end poses are nearest neighbors of π_s and π_e respectively.

$$\hat{\phi}(\pi_s, \pi_e) = \underset{\phi_i \in \mathcal{N}_{\phi(\pi_s,\pi_e)}}{\arg\max}\ \hat{\mathcal{P}}(\phi_i|\pi_s, \pi_e) \tag{4.9}$$

Entropy Estimation: To estimate the entropy of $\mathcal{P}(\phi|\pi_s)$ from the samples $\mathcal{N}_{\phi(\pi_s)}$, we use the resubstitution estimate of entropy [116] as follows

$$\hat{\mathcal{H}}(\phi|\pi_s) = -\frac{1}{|\mathcal{N}_{\phi(\pi_s)}|} \sum_{\phi_i \in \mathcal{N}_{\phi(\pi_s)}} \log \hat{\mathcal{P}}(\phi_i|\pi_s) \tag{4.10}$$

where $\hat{\mathcal{P}}(\phi_i|\pi_s)$ is obtained from Equation (4.7). Under mild conditions, this estimate has been proved to be consistent in the first and second order means [116].

4.4 Applications

In this section, we briefly enumerate applications which benefit from the dynamic information associated with human pose.

4.4.1 Human Motion Prediction from still images

The proposed method can be used to predict future human poses given a start pose or a combination of start and end poses. We represent the future poses in terms of a sequence

Algorithm 3: Algorithm for finding the most probable future motion and the degree of dynamic information in a test pose.

Input: Test pose π_s, database of training poses and the associated action segments

$$\mathscr{D} = \{(\pi, \phi(\pi))\}$$

Output: Most probable action segment $\hat{\phi}(\pi_s)$, amount of dynamic information DDI(π_s)

1. Sample Computation: Compute the set of nearest neighbor poses $\mathscr{N}_{\pi_s}$ of the test pose π_s. Obtain $\mathscr{N}_{\phi(\pi_s)}$, set of action segments associated with the poses in $\mathscr{N}_{\pi_s}$.

2. Conditional Density Estimation: Using the action segments in $\mathscr{N}_{\phi(\pi_s)}$ as samples, obtain the conditional density using non parametric density estimation, as given by Equation (4.7).

$$\hat{\mathscr{P}}(\phi|\pi_s) = c_1 \sum_{\phi_i \in \mathscr{N}_{\phi(\pi_s)}} \Psi(M^{-\frac{1}{2}}(I_d - \Omega_i^\top \Omega\Omega^\top \Omega_i)M^{-\frac{1}{2}})$$

3. Mode Estimation: Obtain the most probable action segment using Equation (4.9).

$$\hat{\phi}(\pi_s) = \underset{\phi_i \in \mathscr{N}_{\phi(\pi_s)}}{\arg\max}\ \hat{\mathscr{P}}(\phi_i|\pi_s)$$

4. DDI Estimation: Compute DDI(π_s) using Equations (4.10) and (4.2).

$$\hat{\mathscr{H}}(\phi|\pi_s) = -\frac{1}{|\mathscr{N}_{\phi(\pi_s)}|} \sum_{\phi_i \in \mathscr{N}_{\phi(\pi_s)}} \log \hat{\mathscr{P}}(\phi_i|\pi_s)$$

$$\text{DDI}(\pi_s) = \exp[-\hat{\mathscr{H}}(\phi|\pi_s)]$$

of images of humans, as shown in Figures 4.6 and 4.7. This output representation is general, and is independent of the application utilizing the estimated future motion. One can easily apply a pose estimation algorithm [7] on our output representation to obtain the 2D or 3D pose of humans, as required by the application.

Predicting future motion given a start pose: In many robotic applications like rehabilitation, surgical gesture assistance and telemanipulation, robots assist humans or manipulate the same object as humans. Such applications are termed "assistance to manipulation" applications. Jarrasse *et al.* has verified that human motion prediction can significantly improve the performance of the robot in these applications [89]. Motion prediction is also useful in detecting gait anomalities in medical applications and analyzing movements in sports videos. Given a test pose, the most probable action segment is obtained, as explained in Section 4.3.2. The poses associated with this action segment represents the predicted future motion. Unlike model-based approaches for motion prediction, the proposed prediction is not restricted to a particular model and can easily incorporate new training data.

Generating realistic human motion trajectories: Creating realistic human motion is an important requirement in applications like humanoid robot design and animation. Given the current pose of the robot π_s and the final pose π_e, such motion can be obtained by finding the most probable action segment starting from π_s and ending at π_e using the proposed method, as explained in Equation 4.4.

4.4.2 Semi-supervised still image action recognition

As explained in Section 4.2, action recognition from still images has recently gained attention in computer vision literature, with applications in action image retrieval and action recognition from personal photo collections, sports images and newspaper images [99, 100]. Most existing methods assume the presence of labeled action images for training. Labeling requires human supervision, and is expensive and time consuming. However, one can easily collect unlabeled images and videos from public databases like Flickr and YouTube. This has lead to the development of semi-supervised algorithms in computer vision [117, 118]. See [119] for an excellent survey of recent efforts on semi-supervised learning. For still image-based action recognition, Cinbis *et al.* [99] developed a semi-supervised method by querying the web to obtain additional training images. However, due the large variation of images in the internet, the additional images used to learn the classifier often differ widely from the test images, leading to lower performance. Another source of training data which is often easy to acquire in applications like surveillance is unlabeled and unsegmented action videos of humans. In this section, we describe how such videos can be utilized for semi-supervised action recognition using the proposed method.

Since the proposed method for predicting action segments do not require activity labels, it can act as a natural way of propagating labels from the labeled training images to the unlabeled video data. For each labeled training pose, we find the most probable action segments from the unlabeled video data, as explained in Section 4.3.2. If the original training poses are discriminative, the retrieved action segments will belong to the same

action. Hence, we add these action segments as additional training samples, thereby increasing the diversity of the training data. Additionally, one could use DDI to propagate labels from just the informative training poses.

4.4.3 Video Summarization

Recent deluge in multimedia content has necessitated the development of algorithms to concisely represent a video. The goal of video summarization is to capture the relevant information in a video using a fixed number of frames called the key-frames. Numerous criteria have been proposed in the literature for selecting the key-frames in a video. Two popular ones are representation and diversity [3]. Representation criteria prefers the selection of key-frames which are similar to the frames in the video. Diversity favors the selection of key-frames which are not redundant. In the case of videos of humans, the amount of dynamic information in a pose has not yet been utilized for summarization. Since frames with high dynamic information convey the motion of the human over a large number of adjacent frames, they are potential candidates for key-frames. Hence, DDI can also be used for key-frame selection in video summarization applications.

4.5 Experiments

We empirically evaluate the proposed method on action datasets of varying complexity, namely the Weizmann activity dataset [120] with clean background and fixed view point, INRIA XMAS (IXMAS) dataset [121] where actors freely change their orientation and the UCF Sports Activity dataset [122] with large changes in scene and view points. To

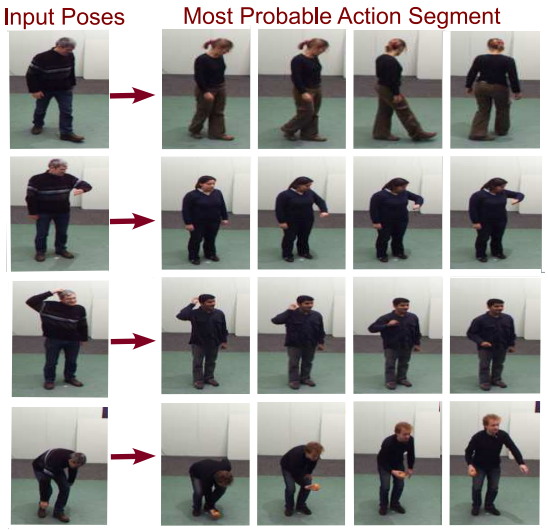


Figure 4.6: Motion prediction using the proposed method. It is interesting to note that the predicted motion is performed by a different subject, since there is no overlap between training and testing subjects.

perceptually evaluate the DDI measure obtained by the proposed method, we predict the amount of dynamic information in the Hokusai Manga images. The fMRI experiments by Osaka *et al.* [2] on these images had demonstrated that the dancer images have higher dynamic information compared to the priest images. Furthermore, to evaluate the method under large variations in training and testing conditions, we perform a cross dataset experiment using unlabeled videos from the Weizmann dataset and test images from the CMU action dataset [123].

4.5.1 Implementation Details

The proposed method is illustrated in Figure 4.5, and details are provided in Algorithm 3. Ballistic boundaries are extracted from unlabeled action videos using [105]. Sequence of

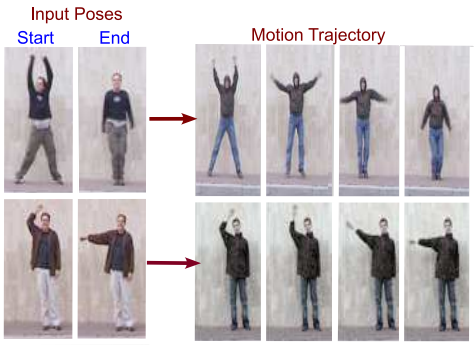


Figure 4.7: Generating trajectories using the proposed method.

poses within adjacent ballistic boundaries form the action segments. A gallery of human poses along with their associated action segments $\mathscr{D} = \{(\pi, \phi(\pi))\}$ is then created from the training videos. The poses are represented by HOG [5] features, and action segments by the finite observability matrix $\Omega_m^\top$ in the LDS model. Closed form expressions exist for the computation of $\Omega_m^\top$ from the action segments, as derived in [111]. Given a test pose π_s, $\mathscr{N}_{\pi_s}$ is created by identifying the k nearest neighbors in the HOG feature space from the gallery. Using the corresponding action segments as samples, mode and entropy of $\mathscr{P}(\phi|\pi)$ are computed as explained in Algorithm 3. Instead of using the iterative optimization algorithm in [115], we compute the most probable action segment directly using (4.9). Unless specified, we fixed k in all our experiments to the average number of repetitions of actions in the unlabeled videos, which is roughly the number of subjects in the unlabeled videos. Our experiments suggest that the proposed method works well over a wide range of k. The bin size and cell size of the HOG features are both set to 8, with 2×2 cells forming a block.

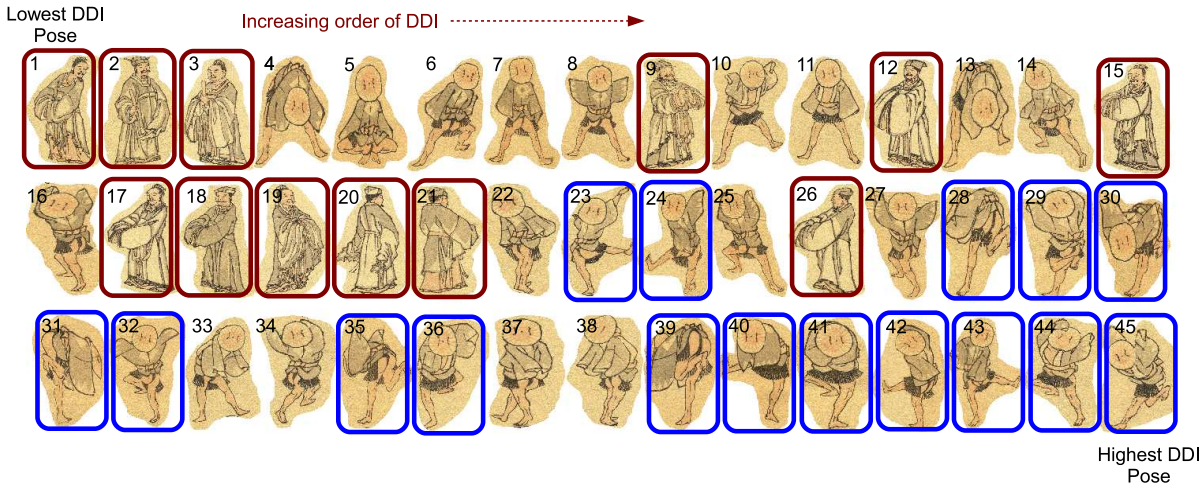


Figure 4.8: The priest and dancer images in the Hokusai Manga collection are displayed in the increasing order of their DDI, with the indices in the sorted order indicated in the top left of each pose. Here, index 1 (top left pose) has the lowest DDI and index 45 (bottom right pose) has the highest DDI. The priest images are marked in red, and the dancer images having the most unstable poses, where the human is standing on a single leg are marked in blue. Observe that most of the priest images have lower DDI values, while most of the dancer images in unstable poses (in blue) have higher DDI values, providing a computational explanation for the results in [2].

4.5.2 Perceptual Evaluation on Manga Images

In this section, we estimate the amount of dynamic information in the Hokusai Manga image database shown in Figure 4.2. This database consists of 45 images belonging to two groups namely the priests and the dancers. The same set of images had been used by Osaka *et al.* [2] in their experiments, which reported that the unstable poses in the dancer images activated the motion sensitive regions of the visual cortex, while the priest images did not. This indicates that the dancer images have higher dynamic information compared

to the priest images.

Since the Manga images have wide variations in human poses, we use the SFU skating dataset [124] for training, as explained in section 4.5.1. For each Manga test image, we do a simple thresholding to obtain a binary image and extract the HOG features. Using the SFU training data, we obtain the DDI values for each Manga image. The Manga images are then sorted in the increasing order of DDI and are displayed in Figure 4.8. The priest images are highlighted in red. As can be observed, most priest images have low DDI values indicating low amounts of implied motion. Furthermore, among the dancer images, the most unstable poses are the ones where the human is standing on one leg. Such images are highlighted in blue. Based on the studies in [2], such unstable poses should have higher implied motion. These images come towards the end of the sorted order in Figure 4.8, indicating that the DDI values are higher in them. Thus, most of the stable poses have lower DDI values, and most of the unstable ones have higher DDI values, there by empirically verifying that the proposed measure is perceptually meaningful.

4.5.3 Human Motion Prediction from still images

We performed motion prediction given a single pose on the IXMAS dataset. We used the first nine subjects in the first view as the training data and predicted the future motion for each pose of the last subject. The predicted motion of some of the test poses are shown in Figure 4.6. It can be observed that the predicted motion mostly agrees with the ones expected by humans.

To evaluate the prediction accuracy, we used the motion prediction error, which is defined

as the difference between the true action segments for each test frame and the predicted action segment. We use the distance metric between action segments defined in (4.6). We plot this error for the proposed method for different values of k in Figure 4.9. The base line method (NN-Based) consists of using the mean of the k retrieved action segments as the predicted motion. Using the first nearest neighbor as the prediction motion, the prediction error is 0.47. The proposed method decreases this prediction error considerably achieving an error of 0.39 using 6 nearest neighbors. Also, the simple baseline of averaging the retrieved nearest neighbor action segments leads to higher prediction error for higher values of k. We attribute the improvement in performance to the following. Due to errors in pose matching, the nearest neighbor poses and their associated motion are often erroneous. These erroneous motion normally form outliers and do not contribute to the most probable motion. Since mean is not robust to outliers, averaging the retrieved action segments lead to poor performance. However, the mode is not sensitive to outliers. Hence, by finding the mode of the nearest neighbor action segments, the proposed method improve the robustness of the motion prediction algorithm to errors in pose matching.

Furthermore, we predicted the probable human trajectories given a start and end pose on the Weizmann dataset. We used two subjects for testing and remaining for training. We illustrate the probable trajectories in Figure 4.7. It is evident that the predicted trajectories are close to the ones expected by humans.

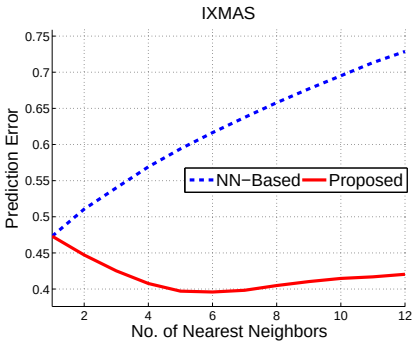


Figure 4.9: Motion prediction error in IXMAS dataset using the nearest neighbor-based(NN-Based) and the proposed method. Due to outliers in the nearest neighbor poses, the NN-Based method lead to lower performance with more nearest neighbors. However, since the proposed method of mode computation is insensitive to outliers, the motion prediction error is reduced with more nearest neighbors by the proposed method.

4.5.4 Semi-Supervised Single Image Action Recognition

In this section, we evaluate the label propagation technique, explained in Section 4.4 for semi-supervised activity recognition from a single image. Since activity recognition from single images arises mainly in sports and newspaper images, we used the UCF Sports Activity dataset in our experiments. We considered nine out of the thirteen actions, avoiding the classes differing only in motion. Action classes which differ only in their motion signature cannot be distinguished in still images, even by humans. Hence, they are not considered for evaluating still image action recognition algorithms in the literature [99], for a fair comparison of the algorithms. The classes used in our experiments are listed in Figure 4.11. We used 2 subjects for training, 2 for testing and 8 as unlabeled data. There is no overlap between the subjects in training, testing and unlabeled data. We chose 8 images at random of the 2 training subjects to form the training data. No labels or temporal

Figure 4.10: For each test image, the nearest neighbors obtained using the supervised method and the proposed method are shown. Erroneous results are encircled in red.

segmentation is assumed for the unlabeled data. We used the HOG features for representing human poses and the nearest neighbor classifier for activity recognition, similar to the approach introduced in [83].

We compared the proposed method of label propagation with the nearest neighbor classifier using the labeled data alone (referred as supervised algorithm) and three popular semi-supervised algorithms namely Self-Training [119], Semi-Supervised SVM (S3VM) [125] and Single View CoTraining [126]. In Self-Training, the classifier trained on the labeled data is applied on the unlabeled data and the L (fixed as 20 in our experiments) most confident images are added to the training set as additional labeled data, using the predicted labels. Test samples are classified using this extended training set. Since S3VM

and Single View CoTraining were originally developed for two class problems, we used one-versus-all classification for multi-class classification. S3VM utilizes unlabeled data by constraining the classifier decision boundary to pass through low density data regions. We used the Multiple Switching algorithm in [125], which iteratively labels the unlabeled data and switches the labels to reduce the optimization cost. Since this algorithm has multiple regularization parameters to be tuned, we compute the recognition accuracy over a wide range of these parameters and report the best results on the test data. The Single View CoTraining algorithm automatically splits the feature vectors into two views, and uses the most confident samples in one view to retrain the other view. It has achieved state of the art results for semi-supervised object recognition [126]. We observed the algorithm to converge in ten iterations and the learned classifier was used for recognition. In the proposed method for label propagation, for each labeled image, we added the k most probable action segments from the unlabeled data into the training set. We used $k = 8$ in our experiments, since unlabeled data contained each action roughly 8 times, performed by each of the 8 subjects. Recognition of test samples were done as before using the extended training set.

The recognition accuracies using 8 action segments are shown in Table 4.1 . We include the corresponding confusion matrices in Figure 4.11. The proposed method provides a significant improvement of 8.6%, compared to the supervised algorithm. Also confusion with wrong classes is considerably reduced. We show some of the test images and the nearest neighbors obtained by the supervised algorithm and the proposed method in Figure 4.10. Furthermore, we plot the variation in accuracy with the number of action segments added in Figure 4.13. As can be observed, the accuracy increases with action

Method	Accuracy (%)
Supervised	49.3
Self-Training	51.9
Semi-Supervised SVM	51.7
Single View CoTraining	53.5
Proposed Method	**57.9**

Table 4.1: Activity Recognition accuracy on the UCF dataset.

segments till 9 (close to 8, the average number of repetitions in the unlabeled data) and then falls gradually.

Additional insight into performance improvement: We attribute the performance improvement of the proposed method to the following. While existing semi-supervised algorithms use unlabeled samples having similar pose as the ones in training, they do not explicitly incorporate the motion information contained in these poses. Since the proposed method models this dynamic information, it is also able to utilize poses in the unlabeled data, which are different from the labeled training samples. This improves the diversity of the training samples and lead to superior performance. Furthermore, the proposed method is insensitive to errors in pose matching.

To gain further insight, we display the nearest neighbor poses and the most probable action segment for two labeled images, belonging to the diving and swing actions re-

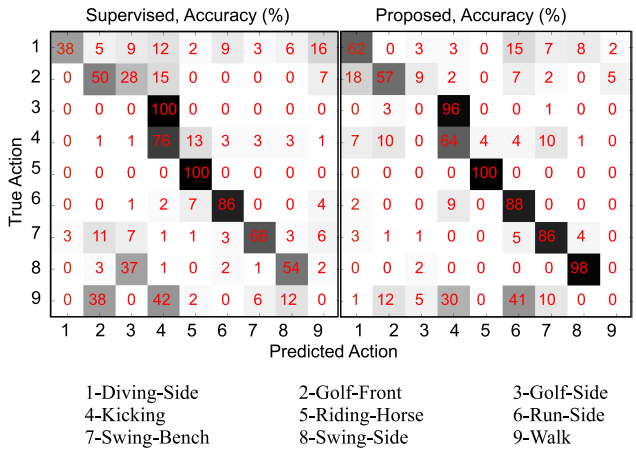


Figure 4.11: Confusion matrices for action recognition on UCF dataset shows significant improvements. In the proposed method, confusion remains mainly between Golf Side and Kicking which have similar leg poses (legs far apart), and among walk, run and kicking, which differ mainly in the rate of execution of the action.

spectively in Figure 4.12. Due to errors in pose matching, some of the nearest neighbors belong to different actions. Since self training adds these samples into training with the wrong label, it corrupts the training data and leads to lower performance. However, since the action segments corresponding to the wrongly retrieved poses differ widely, they usually form outliers during density estimation. The correctly retrieved action segments, whose poses are highlighted in red act as inliers since they are similar to one another. Hence, the most probable action segment belongs to the true action, since it is supported by these inlier action segments. Since the proposed method adds highly probable action segments, it is more robust to pose matching errors. Furthermore, poses in the added action segment are significantly different from the query labeled image, there by increasing the diversity of the training data.

4.5.5 Cross-Dataset Dynamic Inference

In applications like surveillance, the unlabeled and unsegmented videos required for dynamic inference can often be acquired in conditions close to those during testing. However, this is not possible in applications like activity recognition from newspaper images using unlabeled YouTube videos for label propagation. Here, the acquisition conditions of the newspaper images are very different from that of the YouTube videos. To further evaluate the robustness of the proposed method, we consider the scenario where the test pose whose motion is to be inferred, is significantly different from the unlabeled videos available for learning the conditional density. Specifically, we picked poses from the CMU dataset [123] and learned conditional density using the videos from the Weizmann dataset. We then propagated action segments from the Weizmann dataset into the training set as explained before. Test poses in the CMU dataset were recognized using this extended training set. We chose these datasets since they share common actions and differ widely in their acquisition conditions. Out of the four actions in the CMU dataset which are also present in the Weizmann dataset, we use “jumping jack”, “one handed wave” and “pickup” for our experiment. We avoid the fourth action, namely, the “two handed wave”, since it closely resembles jumping jack in still images. Such actions, which cannot be distinguished from still images merely increase the complexity of the still image action recognition problem and make a fair comparison of algorithms difficult. Hence, they are normally removed in the literature [99]. The entire Weizmann dataset is used for learning the conditional density, without assuming any labeling or temporal segmentation. We emphasize that this is a very difficult testing condition due to the large variations

in acquisition conditions between the datasets, heavy clutter in the CMU dataset and the presence of 7 distractor actions in the Weizmann dataset.

Method	Accuracy (%)
Supervised	44.0
Self Training	44.8
Semi-Supervised SVM	45.9
Single View CoTraining	45.2
Proposed Method	**50.5**

Table 4.2: Activity Recognition accuracy on the CMU dataset.

We picked one image per action for training from the CMU dataset and tested on images from the remaining videos. For each training image, we added the most probable action segments for the Weizmann dataset. To reduce the cross-dataset variations, before recognition, we learned a Partial Least Squares(PLS) subspace, using the training samples from the CMU dataset and the added action segments from the Weizmann dataset. PLS-based latent spaces have been effectively used in the literature to handle cross-dataset and cross-model recognition [127]. Interested readers are referred to [127] for further details. We observed the method to be robust to the subspace dimension and chose half the original feature dimension in our experiments. Other evaluation details were similar to those for the action recognition experiment on the UCF dataset before. We present

the recognition accuracies in Table 4.2, and plot the performance with varying number of nearest neighbors (k) in Figure 4.14. The results demonstrate that the proposed method consistently improves the recognition accuracy, even with large variations between the unlabeled gallery and the testing samples.

4.5.6 Video Thumbnailing

We evaluated the proposed dynamic information measure for video thumbnailing. In this problem, the most representative frame in the test video is chosen as the video thumbnail. Our method consists of selecting the image with the highest dynamic information. We compared the proposed scheme with the exemplar selection algorithm called Manifold Précis [3]. We used this method for comparison since it uses the same LDS representation for actions and also achieved state-of-the-art performance. We randomly chose two subjects from the Weizmann dataset for testing and the remaining subjects as the unlabeled gallery. We chose the Weizmann dataset in this experiment, since it consists of short clips of human actions, a setting where thumbnailing becomes a relevant problem. Both structure and motion-based features were used for the Manifold Précis method. HOG features of the pose were used to capture the structure. LDS features computed from a small motion clip centered at each frame were used to capture the motion. It is important to note the proposed method does not use motion information in the test video unlike the Manifold Précis. However, it does require unlabeled and unsegmented videos in the gallery, which are not required in [3].

We show some of the selected key-frames in Figure 4.15. As can be observed, often the

Method	Accuracy(%)
Manifold Précis	85
Proposed Method	**90**

Table 4.3: Nearest neighbor recognition accuracy of the key-frames selected by Manifold Précis and the proposed method.

key frames obtained using the proposed method are more representative of the underlying action. To quantitatively evaluate the two methods, we considered recognizing the action from just the key-frames selected by both the methods. We used images from four subjects, non-overlapping with the test ones to form the training data. The nearest neighbor classifier on HOG features was used for recognition. The recognition accuracies are shown in Table 4.3. The improvements in the obtained accuracy indicate the superior representative properties of the key frames retrieved by the proposed method. In practice, one could combine DDI with existing measures for summarization like representation and diversity [3] to obtain better results.

Figure 4.12: Example illustrating the working of the proposed label propagation approach for semi-supervised action recognition, for two labeled poses in training belonging to the diving and swing actions respectively. The correctly retrieved nearest neighbor poses are highlighted in red. While some of the nearest neighbors belong to incorrect activities due to errors in pose matching, the most probable action segment belongs to the correct class. Furthermore, the poses added by the proposed method are clearly very different from the test pose. Hence, the training set is greatly enriched by the proposed label propagation method.

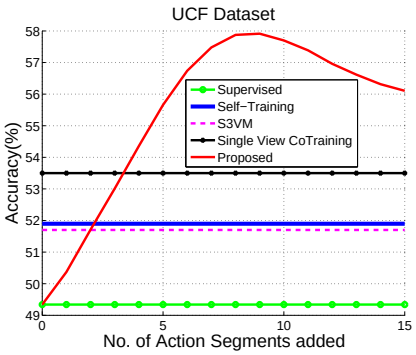


Figure 4.13: Variation of recognition accuracy with the number of action segments added per training image.

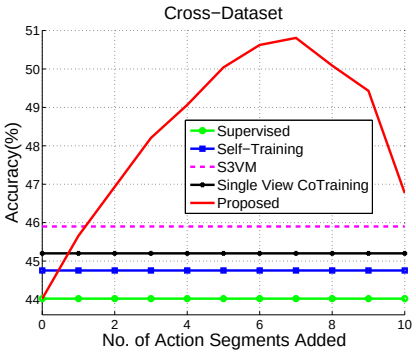


Figure 4.14: Variation of recognition accuracy with the number of nearest neighbors in the CMU cross-dataset experiment.

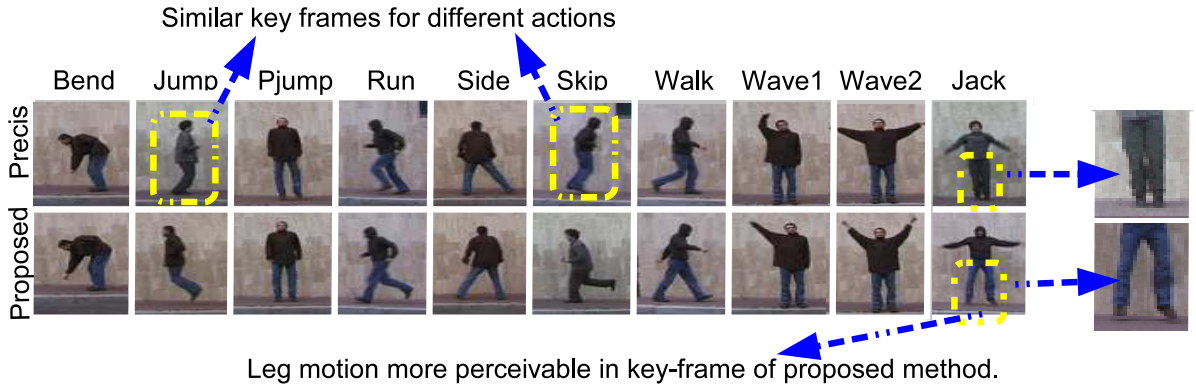


Figure 4.15: Key-frames selected by Manifold Précis [3] and the proposed method. Poses retrieved by [3] for “jump” and “skip” actions are similar. Also motion of legs, which differentiates “jack” from “two handed wave” is more perceivable in the key-frame of the proposed method, as legs are not far apart in the normal standing pose.

Chapter 5

Conclusion and Directions for Future Work

5.1 Summary

In this dissertation, we developed efficient machine learning techniques for visual classification, when labeled data is limited in number. These algorithms used unlabeled data available during testing, or during training and also labeled data in different domain. In particular we discussed three problems namely:

1. Unconstrained iris recognition, where the training data are the clean iris images, which do not capture all the possible variations during testing. Testing samples have large amount of artifacts due to the unconstrained nature of acquisition, but are large in number. Hence we proposed a Sparse Representation based selection and recognition scheme, which learns the underlying structure of clean images. The introduced algorithm simultaneously selects the good iris sectors, recognizes them separately and combines the numerous recognition results using a Bayesian Fusion framework. Furthermore, we demonstrate how to perform video-based recognition and incorporate privacy using Random Projections without affecting the recognition performance.
2. Sensor Adaptation, where most of the enrolled data are using a different sensor than the one used for testing. We provide an efficient solution to this problem, a machine learning technique to adapt iris data collected from one sensor to another.

We provide one of the first solutions to this problem, a kernel learning framework to adapt iris data collected from one sensor to another. Extensive evaluations on iris data from multiple sensors demonstrate that the proposed method leads to considerable improvement in cross sensor recognition accuracy. Furthermore, since the proposed technique requires minimal changes to the iris recognition pipeline, it can easily be incorporated into existing iris recognition systems.

3. Dynamic Inference from human pose, where unlabeled videos are available during training. We utilize these unlabeled videos to extract implicit motion information present in human poses. We pose the inference of this implicit motion information from still images as a non parametric density estimation problem on non-Euclidean manifolds. Statistical inference on the estimated density provide us with quantities of interest like the most probable future motion of a human pose and how informative the given pose is. Our experiments demonstrate that the extracted motion information benefits a variety of applications in computer vision like activity recognition, motion prediction and video summarization.

5.2 Future Work

Several directions of research are possible for the problems and solutions considered in this dissertation. We discuss some of them below.

5.2.1 Semi-supervised algorithms for video-based applications

Visual classification in videos is one of the core problems in computer vision with applications like activity recognition and event classification. In the future, we plan to analyze how unlabeled or weakly labeled data can be utilized for this problem. Labeling in videos is harder due to the inherent ambiguities about the beginning and ending of activities and the need to segment both spatially and temporally. However, often weak labeling information is available with movie and sports videos like scripts, sub titles and audio. Analyzing these weak labeling information and utilizing them for visual classification is a challenging and relevant problem along the lines of the work presented in this dissertation.

5.2.2 Novel cues for video summarization

In the third part of the dissertation, we demonstrated how unlabeled training videos can be used to aid summarization in text videos. This was based on utilizing the inherent motion information in human poses. However, there are other sources of information that can be extracted from an image containing a human like scene properties [128], and the presence and location of objects [84]. The proposed framework can be extended to capture the influence of these sources on the dynamic information conveyed by the human pose, which can in turn lead to novel cues for video summarization.

5.2.3 Weak labeling for 3D Reconstruction

In the future, we plan to utilize unlabeled data for estimating scene geometries in indoor images. While 3D reconstruction from single image is ill posed, one could use strong

prior in indoor scenes like the “Indoor World” model and the appearance and location of furniture [129]. We will explore how such priors can be developed efficiently from unlabeled indoor images.

Bibliography

[1] Elaine M. Newton and P. Jonathon Phillips. Meta-analysis of third-party evaluations of iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 39(1):4–11, 2009.

[2] N. Osaka, D. Matsuyoshi, T. Ikeda, and Osaka M. Implied motion because of instability in hokusai manga activates the human motion-sensitive extrastriate visual cortex: an fmri study of the impact of visual art. *Neuroreport*, 21(4), 2010.

[3] Nitesh Shroff, Pavan Turaga, and Rama Chellappa. Manifold précis: An annealing technique for diverse sampling of manifolds. In *Neural and Information Processing Systems*, 2011.

[4] Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518. IEEE Computer Society, 2001.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[6] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient matching of pictorial structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

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

[8] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *ICCV*, pages 432–439. IEEE Computer Society, 2003.

[9] Jaishanker K. Pillai, Vishal M. Patel, and Rama Chellappa. Sparsity inspired selection and recognition of iris images. In *Proceedings of the 3rd IEEE international conference on Biometrics: Theory, applications and systems*, BTAS'09, pages 184–189, Piscataway, NJ, USA, 2009. IEEE Press.

[10] Jaishanker K. Pillai, Vishal M. Patel, Rama Chellappa, and Nalini K. Ratha. Secure and robust iris recognition using random projections and sparse representations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(9):1877–1893, 2011.

[11] J.K. Pillai, V.M. Patel, R. Chellappa, and N.K. Ratha. Sectored random projections for cancelable iris biometrics. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1838 –1841, 2010.

[12] Kevin W. Bowyer, Karen Hollingsworth, and Patrick J. Flynn. Image understanding for iris biometrics: A survey. *Computer Vision and Image Understanding*, 110:281–307, May 2008.

[13] John Wright, Allen Y. Yang, Arvind Ganesh, Shankar S. Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009.

[14] H. Proena and L. Alexandre. Iris segmentation methodology for non-cooperative recognition. *IEE Proceedings Part-I: Vision, Image and Signal Processing*, 153(2):199–205, April 2006.

[15] Anil K. Jain, Karthik Nandakumar, and Abhishek Nagar. Biometric template security. *EURASIP J. Adv. Sig. Proc.*, 2008, 2008.

[16] Ruud M. Bolle, Jonathan H. Connell, and Nalini K. Ratha. Biometric perils and

patches. *Pattern Recognition*, 35(12):2727–2738, 2002.

[17] Nalini K. Ratha, Jonathan H. Connell, and Ruud M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal*, 40(3):614–634, 2001.

[18] John Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(11):1148–1161, 1993.

[19] R.P. Wildes. Iris recognition: an emerging biometric technology. *Proceedings of the IEEE*, 85(9):1348 –1363, sep 1997.

[20] Yi Chen, Sarat C. Dass, and Anil K. Jain. Localized iris image quality using 2-d wavelets. In Zhang and Jain [130], pages 373–381.

[21] N. D. Kalka, V. Dorairaj, Y. N. Shah, N. A. Schmid, and B. Cukic. Image quality assessment for iris biometric. In *Proceedings of the 241h Annual Meeting of the Gesellschafi flit Klassifikation*, pages 445–452. Springer, 2002.

[22] Hugo Proenca and Luis A. Alexandre. A method for the identification of noisy regions in normalized iris images. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 04*, ICPR '06, pages 405–408, Washington, DC, USA, 2006. IEEE Computer Society.

[23] Xiao-Dong Zhu, Yuan-Ning Liu, Xing Ming, and Qing-liang Cui. A quality evaluation method of iris images sequence based on wavelet coefficients in "region of interest". In *Proceedings of the The Fourth International Conference on Computer and Information Technology*, CIT '04, pages 24–27, Washington, DC, USA, 2004. IEEE Computer Society.

[24] Li Ma, Tieniu Tan, Yunhong Wang, and Dexin Zhang. Personal identification based on iris texture analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12):1519–1533, 2003.

[25] Karen Hollingsworth, Kevin W. Bowyer, and Patrick J. Flynn. Image averaging for improved iris recognition. In Massimo Tistarelli and Mark S. Nixon, editors, *ICB*, volume 5558 of *Lecture Notes in Computer Science*, pages 1112–1121. Springer, 2009.

[26] Y. Du. Using 2d log-gabor spatial filters for iris recognition. *SPIE Biometric Technology for Human Id entification*, 6202, 2006.

[27] Li Ma, Tieniu Tan, Yunhong Wang, and Dexin Zhang. Efficient iris recognition by characterizing key local variations. *IEEE Trans. on Image Processing*, 13:739–750, 2004.

[28] Emine Krichen, Lorène Allano, Sonia Garcia-Salicetti, and Bernadette Dorizzi. Specific texture analysis for iris recognition. In Takeo Kanade, Anil K. Jain, and Nalini K. Ratha, editors, *AVBPA*, volume 3546 of *Lecture Notes in Computer Science*, pages 23–30. Springer, 2005.

[29] Natalia A. Schmid, Manasi V. Ketkar, Harshinder Singh, and Bojan Cukic. Performance analysis of iris-based identification system at the matching score level. *IEEE Transactions on Information Forensics and Security*, 1(2):154–168, 2006.

[30] Chengqiang Liu and Mei Xie. Iris recognition based on dlda. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 489 –492, 0-0 2006.

[31] Kaushik Roy and Prabir Bhattacharya. Iris recognition with support vector machines. In Zhang and Jain [130], pages 486–492.

[32] George I. Davida, Yair Frankel, and Brian J. Matt. On enabling secure applications through off-line biometric identification. In *IEEE Symposium on Security and Privacy*, pages 148–157. IEEE Computer Society, 1998.

[33] Feng Hao, Ross Anderson, and John Daugman. Combining crypto with biometrics effectively. *IEEE Trans. Computers*, 55(9):1081–1088, 2006.

[34] Sanjay Ganesh Kanade, Dijana Petrovska-Delacrétaz, and Bernadette Dorizzi. Cancelable iris biometrics and using error correcting codes to reduce variability in biometric data. In *CVPR*, pages 120–127. IEEE, 2009.

[35] Ari Juels and Martin Wattenberg. A fuzzy commitment scheme. In *ACM Conference on Computers and Communications Security*, pages 28–36. ACM Press, 1999.

[36] Andrew Teoh Beng Jin, Alwyn Goh, and David Ngo Chek Ling. Random multispace quantization as an analytic mechanism for biohashing of biometric and random identity inputs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1892–1901, 2006.

[37] Jinyu Zuo, Nalini K. Ratha, and Jonathan H. Connell. Cancelable iris biometric. In *ICPR*, pages 1–4, 2008.

[38] D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on*, 47(7):2845 –2862, nov 2001.

[39] David L. Donoho and Michael Elad. On the stability of the basis pursuit in the presence of noise. *Signal Process.*, 86(3):511–532, March 2006.

[40] D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):310 –316, april 2010.

[41] Scott Shaobing Chen, David L. Donoho, Michael, and A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

[42] J.K. Pillai, V.M. Patel, and R. Chellappa. Sparsity inspired selection and recognition of iris images. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, pages 1 –6, sept. 2009.

[43] A. Beng Jin Teoh and Chong Tze Yuang. Cancelable biometrics realization with multispace random projections. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 37(5):1096 –1106, oct. 2007.

[44] J.K. Pillai, V.M. Patel, R. Chellappa, and N.K. Ratha. Sectored random projections for cancelable iris biometrics. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1838 –1841, march 2010.

[45] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *Information Theory, IEEE Transactions on*, 54(5):2210 –2219, may 2008.

[46] E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489 – 509, feb. 2006.

[47] David L. Donoho. High-dimensional centrally-symmetric polytopes with neighborliness proportional to dimension. Technical report, Comput. Geometry, (online) Dec, 2005.

[48] Jeffrey D. Blanchard, Coralia Cartis B, Jared Tanner B, and Andrew Thompson B. Phase transitions for greedy sparse approximation algorithms. submitted, 2009.

[49] K. W. Bowyer and P. J. Flynn. The nd-iris-0405 iris image dataset. *Notre Dame CVRL Technical Report*.

[50] P. Jonathon Phillips, Patrick J. Flynn, J. Ross Beveridge, W. Todd Scruggs, Alice J. O'Toole, David Bolme, Kevin W. Bowyer, Bruce A. Draper, Geof H. Givens, Yui Man Lui, Hassan Sahibzada, Joseph A. Scallan, Iii, and Samuel Weimer. Overview of the multiple biometrics grand challenge. In *Proceedings of the Third International Conference on Advances in Biometrics*, ICB '09, pages 705–714, Berlin, Heidelberg, 2009. Springer-Verlag.

[51] Ewout van den Berg and Michael P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, November 2008.

[52] L. Masek and P. Kovesi. Matlab source code for a biometric identification system based on iris patterns. *The University of Western Australia*, 2003.

[53] P Jonathon Phillips Nist. Frgc and ice workshop. *Access*, page 34, 2006.

[54] James R. Matey and Lauren R. Kennell. Iris recognition - beyond one meter. In Massimo Tistarelli, Stan Z. Li, and Rama Chellappa, editors, *Handbook of Remote Biometrics*, Advances in Pattern Recognition, pages 23–59. Springer London, 2009.

[55] Kevin Bowyer, Sarah Baker, Amanda Hentz, Karen Hollingsworth, Tanya Peters, and Patrick Flynn. Factors that degrade the match distribution in iris biometrics. *Identity in the Information Society*, 2:327–343, 2009.

[56] Ryan Connaughton, Amanda Sgroi, Kevin W. Bowyer, and Patrick J. Flynn. A cross-sensor evaluation of three commercial iris cameras for iris biometrics. In *IEEE Computer Society Workshop on Biometrics*, 2011.

[57] A. Ross and A. K. Jain. Biometric sensor interoperability: A case study in fingerprints. In *International ECCV Workshop on Biometric Authentication*, pages 134–145, 2004.

[58] F. Alonso-Fernandez, R.N.J. Veldhuis, A.M. Bazen, J. Fierrez-Aguilar, and J. Ortega-Garcia. Sensor interoperability and fusion in fingerprint verification: A case study using minutiae-and ridge-based matchers. In *International Conference on Control, Automation, Robotics and Vision*, pages 1–6, 2006.

[59] F. Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Gonzalez-Rodriguez. Quality-based conditional processing in multi-biometrics: Application to sensor interoperability. *IEEE Transactions on Systems, Man and Cybernetics*, 40(6):1168–1179, 2010.

[60] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[61] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *Conference on Learning Theory*, pages 144–152, 1992.

[62] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon.

Information-theoretic metric learning. In *International Conference on Machine Learning*, pages 209–216, 2007.

[63] K. Saenko, B. Kulis, M. Fritz, and T.J. Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pages 213–226, 2010.

[64] H. Van Nguyen, V. M Patel, N.M. Nasrabadi, and R. Chellappa. Kernel dictionary learning. In *International Conference on Acoustics, Speech and Signal Processing*, 2012.

[65] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Christopher J. C. H. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

[66] S. V. N. Vishwanathan, Alexander J. Smola, and René Vidal. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *International Journal of Computer Vision*, 73(1):95–119, 2007.

[67] Kilian Q. Weinberger, Fei Sha, and Lawrence K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 106–, New York, NY, USA, 2004. ACM.

[68] John Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1148–1161, 1993.

[69] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[70] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591. MIT Press, 2001.

[71] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[72] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. In *Proceedings of the 21st international jont conference on Artifical intelligence*, IJCAI'09, pages 1187–1192, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[73] Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelliegence*, 26:101–126, 2006.

[74] Brian Kulis, Mtys Sustik, and Inderjit Dhillon. Learning low-rank kernel matrices. In *International Conference on Machine Learning*, pages 505–512, 2006.

[75] L.M. and Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967.

[76] Amanda Sgroi, Kevin Bowyer, and Patrick Flynn. Cross sensor iris recognition competition. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2012.

[77] Yooyoung Lee, R.J. Micheals, and P.J. Phillips. Improvements in video-based automated system for iris recognition (vasir). In *Workshop onMotion and Video Computing*, pages 1–8, 2009.

[78] Yooyoung Lee, R.J. Micheals, and P.J. Phillips. Robust iris recognition baseline

for the grand challenge. In *National Institute of Standards and Technology Interagency/Internal Report*, 2011.

[79] J. E. Prussing. The principal minor test for semidefinite matrices. *Journal of Guidance, Control, and Dynamics*, 9(1):121–122, 1986.

[80] Thomas B. Moeslund, Adrian Hilton, Volker Krüger, and Leonid Sigal, editors. *Visual Analysis of Humans - Looking at People*. Springer, 2011.

[81] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4), 2006.

[82] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3), 2011.

[83] Christian Thurau and Václav Hlavác. Pose primitive based human action recognition in videos or still images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[84] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[85] Lorella Battelli, Patrick Cavanagh, and Ian M. Thornton. Perception of biological motion in parietal patients. *Neuropsychologia*, 41(13), 2003.

[86] Joachim Lange, Karsten Georg, and Markus Lappe. Visual perception of biological motion by form: a template-matching analysis. *Journal of vision*, 6(8), 2006.

[87] Masahiro Hirai and Kazuo Hiraki. The relative importance of spatial versus temporal structure in the perception of biological motion: An event-related potential study. *Cognition*, 99(1), 2006.

[88] Zoe Kourtzi and Nancy Kanwisher. Activation in human mt/mst by static images with implied motion. *Journal of Cognitive Neuroscience*, 12(1), January 2000.

[89] N. Jarrasse, J. Paik, and G. Morel. Can human motion prediction increase transparency? In *International Conference on Robotics and Automation*, 2008.

[90] Sergey Ioffe and David Forsyth. Learning to find pictures of people. In *Neural and Information Processing Systems*, 1998.

[91] Ankur Agarwal and Bill Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1), 2006.

[92] Duan Tran and David Forsyth. Configuration estimates improve pedestrian finding. In *Neural and Information Processing Systems*, 2007.

[93] Qiang Zhu, Shai Avidan, M. Yeh, and K. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[94] Oncel Tuzel, Fatih Porikli, and Peter Meer. Human detection via classification on riemannian manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[95] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 2010.

[96] Deva Ramanan. Learning to parse images of articulated bodies. In *Neural and Information Processing Systems*, 2006.

[97] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE International Conference on Computer Vision*, 2009.

[98] Nazli Ikizler, Ramazan Gokberk Cinbis, Selen Pehlivan, and Pinar Duygulu. Recognizing actions from still images. In *International Conference on Pattern Recognition*, 2008.

[99] Nazli Cinbis, Ramazan Cinbis, and Stan Sclaroff. Learning actions from web. In *IEEE International Conference on Computer Vision*, 2009.

[100] Weilong Yang, Yang Wang, and Greg Mori. Recognizing human actions from still images with latent poses. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[101] Stefan Roth and Michael J. Black. On the spatial statistics of optical flow. *International Journal of Computer Vision*, 74(1), 2007.

[102] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5), 2011.

[103] J. Yuen and A. Torralba. A data-driven approach for event prediction. In *European Conference on Computer Vision*, 2010.

[104] D. Kerzel. A matter of design: No representational momentum without predictability. *Visual Cognition*, 9(1-2), 2002.

[105] S. N. P. Vitaladevuni, V. Kellokumpu, and L. S. Davis. Action recognition using ballistic dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[106] James Hays and Alexei Efros. Im2gps: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[107] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[108] James Hays and Alexei A. Efros. Scene completion using millions of photographs. *Communications of the ACM*, 51(10), 2008.

[109] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, 2003.

[110] Soren Hauberg and Kim Steenstrup Pedersen. Predicting articulated human motion from spatial processes. *International Journal of Computer Vision*, 94(3), 2011.

[111] S. Soatto, G. Doretto, and Y. N. Wu. Dynamic textures. In *IEEE International Conference on Computer Vision*, 2001.

[112] Pavan Turaga, Ashok Veeraraghavan, and Rama Chellappa. Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[113] Yasuko Chikuse. *Statistics on Special Manifolds*. Springer-Verlag, 2003.

[114] O. Tuzel, R. Subbarao, and P. Meer. Simultaneous multiple 3d motion estimation via mode finding on lie groups. In *IEEE International Conference on Computer Vision*, 2005.

[115] E. Cetingul, H. and R. Vidal. Intrinsic mean shift for clustering on stiefel and grass-

mann manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[116] I. Ahmad and P. Lin. A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Trans. on Information Theory*, 22(3), 1976.

[117] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google's Image Search. In *IEEE International Conference on Computer Vision*, 2005.

[118] Lixin Duan, Dong Xu, Ivor Wai-Hung Tsang, and Jiebo Luo. Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9), 2012.

[119] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-supervised learning*. Adaptive computation and machine learning. MIT Press, 2006.

[120] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12), 2007.

[121] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(3), 2006.

[122] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[123] Yan Ke, Rahul Sukthankar, and Martial Hebert. Event detection in crowded videos. In *IEEE International Conference on Computer Vision*, 2007.

[124] Y. Wang, H. Jiang, M/ S. Drew, Z.-N. Li, and G. Mori. Unsupervised discovery of action classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[125] Vikas Sindhwani and S. Sathiya Keerthi. Large scale semi-supervised linear svms. In *International SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[126] Minmin Chen, Kilian Q. Weinberger, and Yixin Chen. Automatic feature decomposition for single view co-training. In *International Conference on Machine Learning*, 2011.

[127] Abhishek Sharma and David W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[128] Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.

[129] Varsha Hedau, Derek Hoiem, and David A. Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, pages 1849–1856. IEEE, 2009.

[130] David Zhang and Anil K. Jain, editors. *Advances in Biometrics, International Conference, ICB 2006, Hong Kong, China, January 5-7, 2006, Proceedings*, volume 3832 of *Lecture Notes in Computer Science*. Springer, 2006.