2015-12-14

Sparse Representation and Dictionary Learning for Biometrics and Object Tracking

Rahman Saeed Khorsandi
University of Miami, saeed.khorsandi@gmail.com

UNIVERSITY OF MIAMI

SPARSE REPRESENTATION AND DICTIONARY LEARNING FOR BIOMETRICS AND OBJECT TRACKING

By

Rahman Khorsandi

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

December 2015

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

SPARSE REPRESENTATION AND DICTIONARY LEARNING FOR BIOMETRICS AND OBJECT TRACKING

Rahman Khorsandi

Approved:

Mohamed Abdel-Mottaleb, Ph.D. Professor of Electrical and Computer Engineering	Shahriar Negahdaripour, Ph.D. Professor of Electrical and Computer Engineering
Kamal Premaratne, Ph.D. Professor of Electrical and Computer Engineering	Mei-Ling Shyu, Ph.D. Professor of Electrical and Computer Engineering
Arun Ross, Ph.D. Professor of Electrical and Computer Engineering Michigan State University	Dean of the Graduate School

KHORSANDI, RAHMAN (Ph.D., Electrical and Computer Engineering)

Sparse Representation and Dictionary Learning for Biometrics and Object Tracking (December 2015)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Mohamed Abdel-Mottaleb.
No. of pages in text. (131)

Biometrics attracted the attention of researchers in computer vision and machine learning for its use in many applications. We propose systems for face and ear recognition, gender classification and object tracking. First, we present a fully automated system for recognition from ear images based upon sparse representation. In sparse representation, extracted features from the training data is used to develop a dictionary. Classification is achieved by representing the extracted features of the test data as a linear combination of entries in the dictionary. In fact, there are many solutions for this problem and the goal is to find the sparsest solution. We use a relatively new algorithm named smoothed l^0 norm to find the sparsest solution and Gabor Wavelet features are used for building the dictionary. Experimental results conducted on the University of Notre Dame (UND) collection J data set, containing large appearance, pose, and lighting variations, resulted in a gender classification rate of 89.49%. Furthermore, the proposed method is evaluated on the WVU data set and classification rates for different view angles are presented. Results show improvement and great robustness in gender classification over existing methods.

Furthermore, we present an approach for gender classification using facial images based upon sparse representation and Basis Pursuit. In sparse representation, the training data is used to develop a dictionary based on extracted features. Basis pursuit

is used to find the best representation by minimizing the l^1 norm. Experimental results are conducted on the FERET data set and obtained results are compared with other works in this area. The results show improvement in gender classification over existing methods.

We present a novel classification technique based on sparse representation. Currently, most of the methods for sparse representation classification do not apply constraints to the coefficients that form the linear combination of the atoms, which leads to having coefficients that can be positive or negative. In addition, all the training samples are treated uniformly without differentiating between the training samples in the dictionary. In this technique, we impose non-negative constraint on the components of the coefficient vector to ensure that the coefficient vector represents the contributions of the training samples towards the query, which is more natural for classification purposes. We also use the mutual information between the query sample and each of the training samples to obtain a weight for each of the atoms in the dictionary. These weights have the effect of reducing the search space and speeding the convergence of the algorithm in finding the coefficient vector. Experiments conducted on the Extended Yale B database for face recognition and on the University of Notre Dame (UND) database for ear recognition show that the proposed non-negative weighted sparse representation obtained by smoothed l^0 norm outperforms other state-of-the-art classifiers.

Finally, a general tracking system is developed based upon sparse representation. Developing an effective and complete tracking algorithm is a challenging task because of factors such as illumination, occlusion and pose variations. Most of the tracking algorithms do not consider the situation when the tracked object or disap-

pears temporarily from the video sequence or becomes temporarily fully occluded. Here, our goal is to develop an automatic object tracking system that can handle pose variations, scale variations and temporary disappearance of the object from the scene. We present a robust tracking system based on adaptive sparse representation and feedback. We focus on automatic tracking with no prior knowledge other than the location of the region to be tracked in the first frame, which can either be located manually or using a detector that finds the region of interest (ROI). The visual tracking is a binary classification problem. The positive samples are bounding boxes that have high overlap with current position of the target while negative samples are drawn from regions outside the ROI to model background close to the target. The tracking algorithm uses the dictionary to locate the ROI in the following frames via adaptive sparse representation. One of the main issues in tracking systems is false tracking when the object disappears from the scene. Motivated by the concept of feedback in control systems, we overcome the problem of false tracking when the object disappears by comparing the newly tracked region with previous regions to confirm that the object is still in the frame. A structural similarity measure is used to measure similarity between a newly tracked ROI and the previously tracked ROIs and if the similarity is below a certain threshold, the object is assumed to be out of the scene. In fact, this similarity evaluation is like a feedback loop in our tracking algorithm which makes our method robust, reliable and accurate when compared to the state-of-the-art methods on challenging sequences. If the object is not located in the current frame, the algorithm stops tracking and starts searching for the object in the following frames. The searching is achieved by using a detector based on

sparse representation and an adaptive dictionary to efficiently locate the object when it reappears in the scene.

to my parents

Acknowledgements

I would like to express my special appreciation to my advisor, Professor Dr. Mohamed Abdel−Mottaleb, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. I would also like to thank my committee members, Dr. Shahriar Negahdaripour, Dr. Kamal Premaratne, Dr. Mei-Ling Shyu and Dr. Aron Ross for serving as my committee members.

Rahman Khorsandi

University of Miami

December 2015

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Sparse representation has received widespread attention because of its robust performance and wide range of applications. During the last decade, the theory of sparse representation has been used in various practical applications in signal processing and pattern recognition [4–11] It has also been used for compression [12], denoising [13], and audio and image analysis [14]. In addition, dictionary learning and sparse representation have been used as powerful tools for recognition, classification and analysis of image and video data [15], [16], [17].

Generally, sparse representation is a technique for reconstructing a signal or image using the fact that signals can be presented by a set of basis elements [18]. To build a robust and efficient recognition system, the number of training samples per subject (or object) is one of the main challenges. Recognition with a single training sample per subject (or object) , unless it is used along with a model, lacks information to predict the variations among different instances of the object. Furthermore, in many applications, several training samples per subject might be available, spanning different variations in illumination, pose or occlusion. In these cases, the features from each sample are extracted and used for the representation and classification of

a query sample. Methods such as the k Nearest Neighbour (kNN) approach, utilize only a subset of the training samples, i. e., the k nearest neighbors, in classification and a query object is assigned a class-label according to the most common class among its k nearest neighbours [19] and kNN places equal weights on all the selected neighbours. However, methods such as Sparse Representation Classification (SRC), that employ the entire training set for decision making have recently shown to significantly outperform the aforementioned methods [20, 21]. Sparse representation uses all the data samples for the decision making and represents the test data as a linear combination of the training data. The assumption is that the set of coefficients of the linear representation can only be represented correctly by the data of one subject. Therefore, ideally the coefficient vector is sparse and the sparseness is used to find the right subject for recognition.

The main approach for obtaining the sparsest solution is based on l^1 norm minimization of the coefficient vector. However, a relatively new approach to find the sparsest solution based on l^0 norm is named SL0, which has been developed mathematically by Mohimani *et al.* [22]. SL0 is a fast algorithm for over complete sparse decomposition. In fact, this method finds sparse solutions for under determined systems of linear equations. Previous methods usually solve sparse problems by minimizing l^1 norm using linear programming (LP) algorithms. However, SL0 algorithm directly minimizes the l^0 norm. SL0 is a fast method which is about two to three orders of magnitude faster than state-of-the-art LP algorithms.

Recently, the sparse representation based classification (SRC) has been successfully used in face recognition. In practical applications, robust face recognition is a challenging task due to the significant variations that can be encountered in face

images. Wright et. al. [20] proposed a face recognition algorithm, based on sparse representation and l^1 norm minimization, which is robust towards variations in lighting conditions, facial expressions and partial occlusions. In fact, the sparse non-zero coefficients should concentrate on the training samples with the same class label as the query sample. On the other hand, due to the rich information contained in ear images, the ear is becoming an important biometrics for recognition and identification. Motivated by the success of sparse representation and reported results in face recognition, in [23] and [24] we presented sparse representation methods for ear recognition and gender classification.

The objective of this thesis is to propose novel systems for face and ear recognition based upon sparse representation. In chapter 2, we describe the sparse representation and feature extraction methods. In chapter 3, a novel approach for ear recognition and gender classification using 2D ear images is proposed. In chapter 4, we present a new system based on sparse representation for gender classification using face images. In chapter 5, Classification based on Weighted Sparse Representation using Smoothed l^0 Norm with Non-negative Coefficients is proposed for face and ear recognition. In chapter 6, we present a robust biometrics recognition system using joint weighted dictionary learning and smoothed L0 norm. In chapter 7, Robust Object Tracking system is presented via adaptive sparse representation and feedback

1.1 Related Work

In recent years, the sparse representation has been widely studied to solve problems in various applications, partially due to the progress of l^0 norm and l^1 norm

minimization techniques [25]. Due to the fact that signals such as audio and images have naturally sparse representations, sparse representation has proven to be a powerful tool for acquiring, representing and compressing signals. In fact, the signals are downsampled using a lower sampling frequency than Shannon-Nyquist rate which leads to efficient estimation, compression and modeling [26], [27], [28]. The idea of sparse representation classification (SRC) for face recognition [20] is to linearly represent a query in terms of all training samples, where most of the coefficients associated with the training samples are zeros as the query sample belongs to only one class. Huang *et al.* [29] sparsely coded a signal on a group of redundant bases and the classification of the signal was performed using the obtained coding vector. Their proposed algorithm includes two terms,the first term measures the signal reconstruction error and the second term measures the sparsity. Geo *et al.* [30] proposed kernel sparse representation for image classification and face recognition. They believed that the use of the kernel trick can capture the non-linear representation of features, which may reduce the feature quantization error and boost the sparse coding performance. Yang *et al.* [9] used Gabor features for SRC with a learned Gabor occlusion dictionary to reduce the computational cost. The number of dictionary columns is reduced in the learned Gabor dictionary. Cheng *et al.* [31] proposed a process to build a directed l^1-graph, in which each sample is represented by a vertex and the ingoing edge weights to each vertex describe its l^1 norm reconstruction from the remaining samples. Yang *et al.* [32] developed an extension of the spatial pyramid matching (SPM) method, by generalizing vector quantization to sparse coding followed by multi-scale spatial max pooling, and proposed a linear SPM kernel based on sparse coding of SIFT descriptors. Patal *et al.* [33] proposed a face recognition algorithm based on dictionary

learning and sparse representation. A dictionary is learned for each class based on the training samples. The representation error is minimized using sparseness constraint and the query sample is projected onto the span of the training data in each learned dictionary. It happens that the generated dictionary has a huge size, which is time consuming. To overcome this disadvantage, Yang *et al.* [34] proposed an unsupervised dictionary learning algorithm to obtain atoms for each class. Lu *et al.* [35] proposed a locality weighted sparse representation based classification method which utilizes both data locality and linearity. Xu and Yang [19] presented a technique that combines sparse representation with the theory of fuzzy sets. They imposed the constraint of non-negative coefficients on the sparse representation and then constructed the fuzzy class membership matrix to assign the membership of the query sample to each class.

There are several possibilities to improve the sparse representation through the dictionary learning, feature extraction, feature fusion and the optimization procedure. In this thesis, we use a fast optimization algorithm to find the l^0 norm. Direct minimization of l^0 norm is difficult because of the fact that the l^0 norm of a vector is a discontinuous function of that vector. An efficient algorithm was proposed in [22] for sparse decomposition based on the smoothed l^0 norm, which is about two to three orders of magnitude faster than the interior-point Linear Programming (LP) solvers, without sacrificing the accuracy. The basic idea of SL0 is to approximate the discontinuous l^0 norm by a continuous function before optimization. The SL0 algorithm can avoid being early trapped at a local extremum. This algorithm was used in various applications such as blind source separation [36], ear recognition [23], [37], decomposition of EEG signals [38] and image super-resolution [39].

Over the past decades, biometrics technologies, such as face or fingerprint recognition, have been used in different applications including security, psychology and human-computer interaction. However, ear biometrics is a relatively new area of research. Similar to face recognition, ear recognition is passive, *i.e.*, an individual can be scanned and their identity confirmed, without the subject actively engaging the device, compared to fingerprints. In addition, ears may be more reliable than faces, since subjects can change their facial expression or manipulate their visage. However, it is not easy to change the ear shape. One of the first approaches in ear biometrics was proposed by Burge and Burger [40]. They used graph-matching techniques on a Voronoi diagram of curve segments extracted from a Canny edge map. Another method was presented by Hurley *et al.* [41] for performing ear recognition, where they represented each ear image by a set of wells and channels. They assumed that for each subject, the ear image contains a set of unique wells and channels which can be used for subject recognition. Abdel-Mottaleb and Zhou [42] proposed an approach for ear recognition from profile images of the face. In their method, ridges and ravines are extracted for ear representation. Alignment between a probe and a gallery model is performed using Partial Hausdorff Distance. The reported results for face recognition via sparse representation are encouraging enough to extend the sparse representation algorithm to other biometrics such as ear biometrics and further evaluate their performance under the most challenging practical conditions such as pose variations and occlusion. One of first approaches for ear biometrics using sparse representation was proposed by Naseem et al. [3]. They used l^1 norm minimization to find the sparsest solution. They conducted several experiments using the University of Notre Dame (UND) database. Kumar and Chan [43] proposed an approach for ear recognition and

verification using sparse representation of local gray-level orientations. The aforementioned methods and papers motivated us to explore a more robust approach for face and ear recognition.

CHAPTER 2

Sparse Representation

2.1 Overview

Theoretical developments of sparse signal representation have been interesting for researchers to use this powerful tool for computer vision and machine learning applications. Over the past decades, there have been many fundamental progress in the field of machine learning. However, there are problems in dealing and processing of the high-dimensional data. During the last decade, a significant research effort has been devoted to find the compact or sparse representation for signals in order to process the large-scale data. Based on sparse representation theory, a signal can be decomposed into a linear combination of a few basic signals which is capable of representing the majority information conveyed by the target signal.

In fact, a sparse signal can be represented as a linear combination of a relatively few base elements in an over complete dictionary. To find sparse representations, we need to solve an under determined system of linear equations for sparsest solution. Sparse representation has recently found various applications in practical areas of signal

processing and pattern recognition [8], [5], [6], [7], [9]. Sparse signal representation has been used for compression, denoising and analysis of audio and image data.

This thesis proposes an investigation into biometrics which exploits sparse representation for improving face and ear recognition algorithms. In this chapter, we review the theories and algorithms that form the basis of sparse representation. It gives an overview of sparse representation and its mathematical development. First, we explain how to build a dictionary. Then sparse representation is formulated and l^0 norm minimization is described. Finally, we introduce the concept of the smooth l^0 norm and explain the SL0 algorithm.

2.2 Building the Dictionary

Sparse representation has been interesting for researchers in signal and image processing since many natural signals have a sparse or compressible representation in a variety of domains, such as Wavelet, discrete Sine transform (DST), discrete cosine transform (DCT) or Fourier domain. A sparse signal refers to a signal which admits a transform domain representation and most coefficients are zero. In other words, a sparse signal can be represented as a linear combination of a relatively few base elements in an over complete dictionary. As a matter of fact, sparse representation introduces a precise mathematical framework to process high-dimensional data that a few coefficients can represent the majority information from the target signals.

The first step in using sparse representation is to build a dictionary using the training data. The dictionary is a matrix in which each column is the feature vector of one of the training samples. Suppose we have a signal vector (or extracted feature

vector from an image) v which for simplicity assumed to be real, i.e. $v \in \mathbf{R}^m$. Assume that there are n_i training data samples for the i^{th} class, where each data sample is represented by a vector of m elements. These vectors are then used to construct the columns of matrix $\mathbf{A}_i$:

$$\mathbf{A}_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, ..., \mathbf{v}_{i,n_i}] \in \mathbf{R}^{m \times n_i} \tag{2.1}$$

$\mathbf{v}_{i,j}$, where $j = 1, ..., n_i$, is a column vector that represents the features extracted from the training data sample j of subject i. Concatenating the matrices $\mathbf{A_i}, i = 1, 2, ..., k$ yields:

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_k] \in \mathbf{R}^{m \times n} \tag{2.2}$$

where k is the number of subjects and $n = \sum_{i=1}^{k} n_i$. In fact, A is the main dictionary which obtained using all training samples from database.

2.3 Sparse Representation

In the theory of sparse representation, it is assumed that a feature vector of a test data from class i can be represented as a linear combination of the feature vectors of the training data from that class [20]:

$$\mathbf{y} = \alpha_{i,1}\mathbf{v}_{i,1} + \alpha_{i,2}\mathbf{v}_{i,2} + ... + \alpha_{i,n_i}\mathbf{v}_{i,n_i} \tag{2.3}$$

where $\mathbf{y} \in \mathbf{R^m}$ is the feature vector of the test data and the $\alpha_{i,j}$ values are the coefficients corresponding to the training data samples of subject i. A linear representation for the feature vector of the test data, $\mathbf{y}$, can then be given as:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbf{R}^m \tag{2.4}$$

where $\mathbf{x}$ is the coefficient vector. Any $\mathbf{x}$ solving this system of equations gives a representation of $\mathbf{y}$. Since A is a m by n, with m is more than 1, there are infinitely many such bfx's as the above system is under determined. By solving this equation for $\mathbf{x}$, the class of the test data $\mathbf{y}$ can be identified. Note that in equation all the training data samples of a given subject are used to form a representation of the test data.

2.4 l^0 Norm Minimization

To solve the system of equations $\mathbf{y} = \mathbf{Ax}$, the number of equations, m, and unknown parameters, n, are important. If $m = n$, the system of equations will be complete and the solution will be unique. However, in recognition and classification, usually there are many subjects or classes, where the test image belongs to only one of the classes and does not belong to the other classes. In addition, the number of extracted features are much less than the training samples. Therefore, the number of equations is less than the number of unknown parameters ($m < n$) and there is no unique solution for the system $\mathbf{y} = \mathbf{Ax}$.

In this approach, the matrix A is the dictionary that contains the representations of n samples or atoms, where each sample is represented by a feature vector of length m. Since $m < n$, it is an over complete matrix. Since dictionary A contains redundancies, it is possible to find $\mathbf{x}$ in an infinite number of ways. Therefore, it is important to introduce a criterion in order to find the best representation. When the system $\mathbf{y} = \mathbf{Ax}$ is under determined (*i.e.*, $m < n$), usually the l^2 norm is used and

the estimate is expressed as follows:

$$(l^2): \quad \widehat{\mathbf{x}}_2 = argmin\|\mathbf{x}\|_2 \quad Subject\ to \quad \mathbf{y} = \mathbf{A}\mathbf{x} \tag{2.5}$$

where $\widehat{\mathbf{x}}_2$ is the solution, which can be obtained simply by computing the pseudo-inverse of $\mathbf{A}$. However, this solution is not useful in all applications. A fresh point of view have been suggested for these under determined systems to solve the problem of indeterminacy. In fact, we should look for the sparsest solution of all the solutions. To measure the sparsity, the l^0 norm is defined as the number of nonzero elements of a vector,

$$\|\mathbf{x}\|_0 = \#\{j|x_j \neq 0\} \tag{2.6}$$

Vector $\mathbf{x}$ is called sparse when $\|\mathbf{x}\|_0 \ll n$ for $\mathbf{x} \in R^n$.

In classification, the test data can only belong to one of the classes represented in the dictionary. Therefore, the obtained answer should be sparse (only a few elements are not zero). The sparsest solution of $\mathbf{y} = \mathbf{A}\mathbf{x}$ can be obtained by minimizing l^0 norm. For this solution, the number of non-zero elements should be minimized as follows [20]:

$$(l^0): \quad \widehat{\mathbf{x}}_0 = argmin\|\mathbf{x}\|_0 \quad Subject\ to \quad \mathbf{y} = \mathbf{A}\mathbf{x} \tag{2.7}$$

where $\|.\|_0$ is the zero norm, which counts the number of non zero elements of $\mathbf{x}$. In general, this optimization problem is known as NP-hard. In fact, we are trying to find the solution with least number of nonzero, among all solutions for an under determined system. However, finding the minimum l^0 norm is not an easy task and does not have a practical procedure for finding the sparsest solution since it is a discontinuous function and we can not use normal optimization algorithms.

In addition, it will be harder as the feature space dimension or number of training samples increase since we need to use combinatorial search. Furthermore, noise affects the solution because noise magnitude can significantly change the l^0 norm of a vector.

Because of all those obstacles, a huge work is dedicated by researchers to find a reasonable method for solving NP-hard problems. There are methods to obtain the sparsest solution of the equation $\mathbf{y} = \mathbf{Ax}$ without dealing with l^0 norm. In this thesis, we use l^1 norm minimization to find the sparsest solution. Furthermore, we use the smoothed l^0 (SL0) algorithm which approximates the l^0 by a continuous function and uses convex optimization to obtain the optimal solution. To best of our knowledge, it is the first work which uses the SL0 algorithm for classification and recognition.

To obtain the sparsest solution and address the computational issue, the l^1 norm optimization is introduced as:

$$(l^1): \quad \widehat{\mathbf{x}}_1 = argmin\|\mathbf{x}\|_1 \quad Subject\ to \quad \mathbf{y} = \mathbf{Ax} \tag{2.8}$$

where $\|\mathbf{x}\|_1$ denotes the l^1 norm of $\mathbf{x}$. The optimization of this equation is not NP-hard problem and and the solution can be found in linear time. It is proven that if the solution of l^0 is sufficiently sparse, it is identical with the solution of l^1 norm [44].

In order to deal with noise, l^1 minimization problem can be extended to the following problem:

$$(l^1): \quad \widehat{\mathbf{x}}_1 = argmin\|\mathbf{x}\|_1 \quad Subject\ to \quad |\ \mathbf{y} - \mathbf{Ax}\ | \leq \varepsilon \tag{2.9}$$

where $\varepsilon > 0$ is a given tolerance.

2.5 Smoothed l^0 Norm (SL0) Algorithm

In order to find the sparsest solution, we use an approach to solve the system $\mathbf{y} = \mathbf{Ax}$ based on the smoothed l^0 norm (SL0) [22]. This algorithm is used to obtain sparse solutions of under determined systems for linear equations by minimizing the l^0 norm. SL0 method is more efficient than the l^0 and l^1-norms in term of computational complexity [22].

The l^0 norm of a vector is a discontinuous function and therefore it is highly sensitive to noise. In addition, combinatorial search is needed for minimizing l^0 which is time consuming. The idea of SL0 is based on the approximation of the discontinuous l^0 norm function using a continuous one such as Gaussian. This approximation is performed using a parameter (σ) which determines the quality of the approximation. Once we obtain a continuous function, it is possible to use convex optimization methods, such as LevenbergMarquardt, GaussNewton or gradient descent for minimization [45].

One example for such approximations is as follows:

$$f_\sigma(x) \triangleq exp(\frac{-x^2}{2\sigma^2}) \tag{2.10}$$

And approximately:

$$f_\sigma(x) \approx \begin{cases} 1, & if \quad |x| \ll \sigma \\ 0, & if \quad |x| \gg \sigma \end{cases}$$

Then, the idea is to minimize l^0 norm, $\|\mathbf{x}\|_0$, using the following function:

$$F_\sigma(\mathbf{x}) = \sum_{i=1}^{r} f_\sigma(x_i) \tag{2.11}$$

In recognition problems, r is the total number of training samples. Hence, it is obvious that for small values of σ, $\|\mathbf{x}\|_0 \approx r - F_\sigma(\mathbf{x})$ and to find the minimum l^0 norm solution, $F_\sigma(\mathbf{x})$ should be maximized.

There is an important issue about the value of σ is in this method. Very small values of σ will result in a nonsmooth $F_\sigma(\mathbf{x})$ and many local maxima and the results will not be accurate. On the other hand, for large values of σ, the results for maximizing the $F_\sigma(\mathbf{x})$ will be similar to that of the l^2 norm [22].

Briefly, SL0 algorithm tries to maximize $F_\sigma(\mathbf{x}) \triangleq \sum_i \exp(-\mathbf{x}_i^2/2\sigma^2)$ for a given value of σ subject to $\mathbf{y} = \mathbf{A}\mathbf{x}$. A decreasing sequence of σ is defined to improve the performance and decrease the chances of trapping in local extrema. For the initial value of σ, F_σ is maximized subject to $\mathbf{y} = \mathbf{A}\mathbf{x}$ using the steepest ascent approach. The $\mathbf{x}$ that maximizes F_σ will be the starting point to find $\mathbf{x}$ that maximizes F_σ for the next (smaller) σ.

In steepest ascent approach, each iteration moves in the desired direction ($\mathbf{x}' \leftarrow \mathbf{x} + \eta \nabla F_\sigma$), followed by projection to the feasible set $\mathcal{S} = \{\mathbf{x}|\mathbf{y} = \mathbf{A}\mathbf{x}\}$ [46]:

$$\widehat{\mathbf{x}}_0 = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{x}'\| \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x} \tag{2.12}$$

$$= \quad \mathbf{x}' - \mathbf{A}^\dagger(\mathbf{A}\mathbf{x}' - \mathbf{y})$$

where $\mathbf{A}^\dagger = \mathbf{A}^\mathbf{T}(\mathbf{A}\mathbf{A}^\mathbf{T})^{-\mathbf{1}}$ is the pseudo-inverse of $\mathbf{A}$. Moreover, $\mathbf{x}$ is initialized by the minimum l^2 norm solution of $\mathbf{y} = \mathbf{A}\mathbf{x}$, that is, $\mathbf{A}^\dagger\mathbf{y}$.

SL0 algorithm as discussed in [22] has three main steps:

1. Initialization: Obtain $\mathbf{x}'_\mathbf{0}$, the solution that minimizes the l^2 norm of $\mathbf{y} = \mathbf{A}\mathbf{x}$ using pseudo-inverse of $\mathbf{A}$ ($\mathbf{A}^\dagger$).

Also, choose a decreasing sequence for σ, $[\sigma_1 \dots \sigma_K]$.

2. For $k = 1, \dots, K$:

   - Initialization: $\mathbf{x} = \mathbf{x'}_{k-1}$
   - For $\ell = 1 \dots L$ (loop L times):
     (a) Let, $\bigtriangledown_{\mathbf{x}} \mathbf{F}_\sigma(\mathbf{x}) \propto -[x_1 exp(-x_1^2/2\sigma^2) \quad ... \quad x_n exp(-x_n^2/2\sigma^2)]^T$
     (b) Let $\mathbf{x} \leftarrow \mathbf{x} - \eta \bigtriangledown_{\mathbf{x}} \mathbf{F}_\sigma(\mathbf{x})$ (η is a small positive constant)
     (c) Project $\mathbf{x}$ back onto the feasible set:

$$\mathbf{x} \leftarrow \mathbf{x} - \mathbf{A}^\dagger(\mathbf{A}\mathbf{x} - \mathbf{y})$$

   - Set $\mathbf{x}'_k = \mathbf{x}$.

3. Finally, the sparsest solution is $\widehat{\mathbf{x}}_0 = \mathbf{x}'_K$

2.6 Conclusion

In this chapter, we briefly explained the concept of the sparse representation and l^0 norm minimization. In addition, l^1 norm minimization and smoothed l^0 norm are discussed as we use them to obtain the sparsest solution for a system of linear equations. We will use sparse representation and these solutions for recognition and classification purposes in our experiments in next chapters.

CHAPTER 3

Ear Recognition and Gender Classification based upon Sparse Representation

3.1 Overview

Due to the rich information contained in ear images, ear is becoming one of the most important biometrics for recognition and identification. Despite its potential, the ear recognition algorithm developed in recent years are premature for deployment in real-world applications, preforming well only under constrained conditions employing highly cooperative subjects. This results from the challenges associated with effects of illumination variability and pose. We will investigate the effect of different viewing angle on ear recognition and gender classification.

Furthermore, automatic gender identification is useful for applications such as security surveillance [47] and gathering statistics about customers in places such as movie theaters, building entrances and restaurants [48]. Gender classification is performed based on human characteristics such as facial features [49], [50], voice [51] and body movement or gait [52]. Although human face provides information about the gender and age, it is not robust because it is affected by emotions and facial

expressions [53]. However, the appearance of the ear is relatively constant and it has distinctive shape properties.

Among the set of physiological traits for viable biometric use, the ear possesses certain inherent characteristics, such as its stable structure and size, that makes its use advantageous [54]. Although researchers, in recent years, worked on ear biometrics [55], [53], [56], [57], [58], it is the first time to use ear images for gender classification using sparse representation and our results show that it is feasible for this purpose [24].

In this chapter we propose a novel approach for gender classification using ear images and sparse representation obtained with the SL0 algorithm. Experiments and results are presented that include the effect of changing the viewing angle and we also use majority voting as a new method for making the decision in sparse representation. Our experiments show that the method is robust for the change of the viewing angle and the results using majority voting are better than results based on other methods based on sparse representation.

This chapter is organized as follows: Section II presents related work in ear biometrics, sparse representation and gender classification. In Section III, we present a brief mathematical explanation of the sparse representation concept and the proposed method to obtain the sparsest solution. Section IV provides details of the feature extraction and dimensionality reduction. Section V presents experimental results to demonstrate the performance of the proposed method. Conclusions and future research directions are discussed in Section VI.

3.2 Related Work

Over the past decades, biometrics technologies, such as face ore fingerprint recognition, have been used in different applications including security, psychology and human-computer interaction. However, ear biometrics is a relatively new area of research. Same as face, ear biometric is passive, meaning an individual can be scanned and their identity confirmed, without the subject actively engaging the device, as is required for fingerprints. In addition, ears may be much more reliable than a face, which is prone to erroneous identification because of ability of a subject to change their facial expression or otherwise manipulate their visage. On the other hand, in many social interactions, it is important to recognize the gender. Although, there are several studies in gender classification based on facial images, to the best of our knowledge, the proposed approach. In this section, we review the literature on gender classification, ear biometrics and classification based upon sparse representation.

Most of the published papers in gender classification are based on facial images, yet there are some studies based on voice [51] and Gait [52]. Moghaddam et al. [59] used Support Vector Machines (SVMs) for gender classification from facial images. They used low resolution thumbnail face images (21×12 pixels) obtained from FERET[1] database. Wu et. al. [60] presented a real time gender classification system using Look-Up-Table Adaboost algorithm. They extracted demographic information from human faces. Gollomb *et al.* [3] developed a neural network based gender identification system [49]. They used face images with resolution of 30x30 pixels from 45 males and 45 females to train a fully connected two-layer neural network, SEXNET. Also,

[1]http://www.nist.gov/humanid/colorferet/home.html

Cottrell and Metcalfe [61] used neural networks for the classification of emotions and gender from facial images. Gutta and Wechsler [62] used hybrid classifiers for gender identification from facial images. The authors proposed a hybrid approach that consists of an ensemble of RBF neural networks and inductive decision trees. Yu et al. [52] presented a study and analysis of gender classification based on human gait. They also used psychological experiments to improve classification accuracy. They used model-based gait features such as height, frequency and angle between the thighs. Gnanasivam et al. [63] compared several classifiers such as Bayes classifier, K-Nearest Neighbour and neural network for gender classification using ear images and the best result was obtained using KNN. In their experiments, they used their own data which is not publicity available. Because of not enough information about their method, specially preprocessing method, it is not possible to compare our results with their method. Instead, we compare our method based on sparse representation with KNN classifier.

Using ear data is relatively a new area in identification and recognition. One of the first approaches in ear biometrics was proposed by Burge and Burger [40]. They used graph-matching techniques on a Voronoi diagram of curve segments extracted from a Canny edge map. Yan and Bowyer [64] proposed two recognition systems based on PCA (eigen-ear) and ICP matching. They used manually labelled ear landmarks to crop the ear region in each range image. Then, located landmarks on the Triangular Fossa and Incisure Intertragica were used for alignment of the ear images. Also, landmarks were used to align the range images for the ICP-based method. Recently, Zhou et al. [65] proposed a recognition system using local and global features, *i.e.*, SPHIS for local key point representation and fixed voxelization for global representation.

Also, Bustard and Nixon [66] proposed an approach which focuses on ear images, where the ear region is treated as a planar surface that is registered to a gallery using a homography transform. In fact, the ear is registered to a gallery using a homography transform calculated using scale-invariant feature-transform (SIFT). They claim that the feature matches reduce the gallery size and enable a precise ranking using a simple distance algorithm. Kumar et al. [43] used local Radon transform for representing the shape of the ear. Also, they investigated the effectiveness of local curvature encoding using Hessian based feature representation. Patal *et al.* [33] proposed a face recognition algorithm based on dictionary learning and sparse representation. A dictionary is learned for each class based on given training samples. The representation error is minimized using sparseness constraint. The test sample is projected onto the span of the training data in each learned dictionary.

The main idea of using sparse representation for recognition is to represent the test data as a linear combination of training data. The set of coefficients of the linear representation is called coefficient vector. If we assume that there are many subjects in database, the test data will only be related to one of the subjects. Therefore, ideally the coefficient vector should be sparse and it is important to find the sparsest solution. In this chapter , we use SL0 algorithm to find the sparsest solution for classification. SL0 is a fast algorithm for over complete sparse decomposition. In fact, this method finds sparse solutions for under determined systems of linear equations. Previous methods usually solve sparse problems by minimizing l^1 norm using linear programming (LP) algorithms.

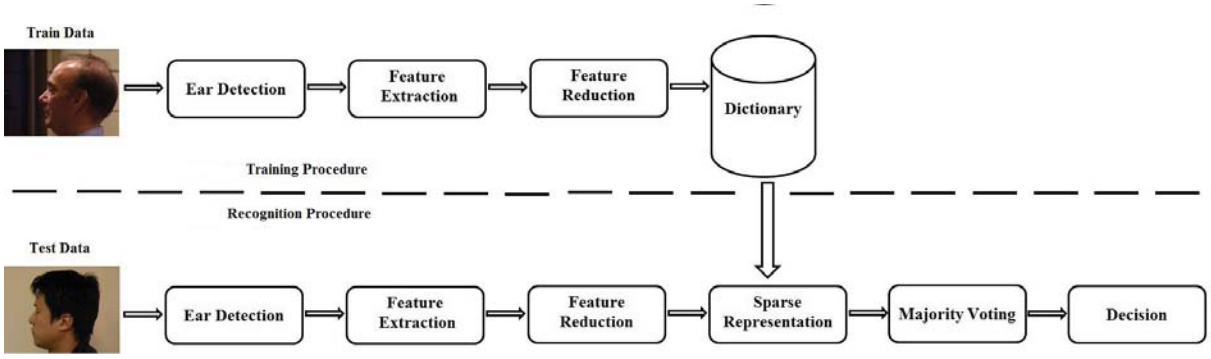


Figure 3.1: System Block Diagram: Gender Classification Using Ear Images

3.3 Proposed Algorithm

In this chapter, we present a system for gender classification from 2-D ear images using sparse representation. Fig. 7.1 shows the block diagram of the proposed system. During training, training samples are used to build a dictionary. Some steps are common between the training and the classification procedures such as ear detection, feature extraction and feature dimension reduction.

Following is a brief explanation of the different steps:

- Ear Detection: given a profile view image, the ear part of the image is localized and a rectangular boundary around the ear region is output using our method described in [57]. A few samples of detected ears are shown in Fig. 3.2.

- Feature Extraction: for each detected ear region, a feature vector is extracted using Gabor wavelets.

Figure 3.2: *Ear Samples frome UND database: Up) Males, Down) Females*

- Feature dimensionality reduction: in order to decrease the computational complexity, the dimensionality of the feature vector is reduced to a lower dimension using Principal Component Analysis (PCA).

- Sparse representation: the basic idea is to represent a test image as a sparse linear combination of the entire training set. Sparse representation will be discussed in the next section.

- Majority Voting: the coefficients obtained using sparse representation are used to perform gender classification. In this part, instead of using the conventional classification using obtained coefficients (sparsest solution), we proposed a new method for classification based on majority voting.

3.4 Gabor Wavelets

The Gabor wavelets (kernels) with orientation μ and scale ν are defined as [1]

$$\psi_{\mu,\nu} = \frac{\| k_{\mu,\nu} \|^2}{s^2} e^{(\frac{-\|k_{\mu,\nu}\|^2 \|z\|^2}{2s^2})} \left[e^{ik_{\mu,\nu}z} - e^{-s^2/2} \right] \tag{3.1}$$

where $z = (x, y)$ is the pixel position, and the wave vector $k_{\mu,\nu}$ is defined as $k_{\mu,\nu} = k_\nu e^{i\phi_\mu}$ with $k_\nu = k_{max}/f^\nu$ and $\Phi_\mu = \pi\mu/8$. k_{max} is the maximum frequency, and f is the spacing factor between kernels in the frequency domain. The ratio of the Gaussian window width to wavelength is determined by s. Considering Eq. 3.1, the Gabor kernels can be generated from one wavelet, *i.e.*, the mother wavelet, by scaling and rotation via the wave vector $k_{\mu,\nu}$ [67]. In this work, we used five scales, $\nu \in 0, ..., 4$ and eight orientations $\mu \in 0, ..., 7$. We also used $s = 2\pi$, $k_{max} = \pi/2$ and $f = \sqrt{2}$.

3.4.1 Feature Extraction using Gabor Wavelets

Image features are extracted using Gabor wavelets by the convolution of a 2-D image with a family of Gabor wavelets as follows [67]:

$$C_{\mu,\nu}(z) = I(z) * \psi_{\mu,\nu}(z) \tag{3.2}$$

Where $z = (x, y)$, $I(z)$ is the 2-D image, and $C_{\mu,\nu}(z)$ is the convolution output at orientation μ and scale ν.

The feature vector φ is constructed out of the $C_{\mu,\nu}(z)$ by concatenating its rows (or columns). Let $C^\beta_{\mu,\nu}$ be the normalized and down sampled (by a constant value β) vectors constructed from $C_{\mu,\nu}(z)$. The Gabor feature vector φ is as follows:

$$\varphi = \begin{pmatrix} C^{\beta^t}_{0,0} & C^{\beta^t}_{0,1} & ... & C^{\beta^t}_{4,7} \end{pmatrix}^t \tag{3.3}$$

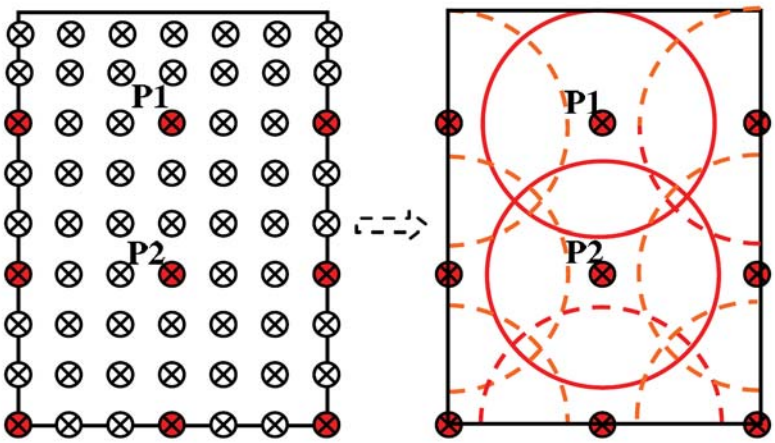


Figure 3.3: *The Uniform Down-Sampling of Gabor wavelets [1]*

where t is the transpose operator.

φ is used as the feature vector in the dictionary matrix used for the sparse representation. Uniform down-sampling by a constant value β to select Gabor feature vector φ is shown in Fig. 3.3 [1]. Red circles show the selected features. Also, one example of extracted features of one image for 8 orientations and 5 spatial frequencies is shown in Fig. 3.4.

3.5 Experiments and Results

In this section, we describe the experiments that we performed in order to evaluate the proposed approach and present the results. We also present comparisons of our classification approach with the well known NN and NS classifiers. We present experiments for recognition and gender classification using two different databases as described in the following sections.

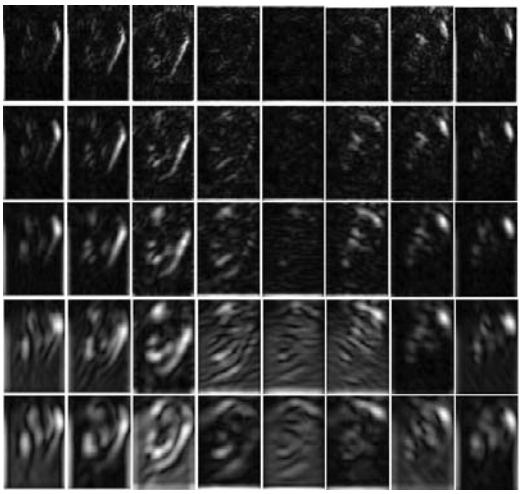


Figure 3.4: *Gabor Features for 5 scales and 8 orientations*

3.5.1 Databases

Experiments are performed on two different data sets. The first data set is the University of Notre Dame (UND) data set collection J2 for 415 subjects which contains profile face images. A few sample ear images for both male and female subjects are shown in Fig. 3.2. The ear region is extracted in each image automatically using our algorithm in [57] which uses a shape based feature set, termed the Histogram of Indexed Shapes (HIS), to localize a rectangular region that contains the ear region. The second data set is the WVU data set, which consists of video sequences captured by a rotating camera around the head of different subjects.

3.5.2 Ear Recognition Experiments

The University of Notre Dame (UND) data set is used to validate the proposed method. A shape-based feature set, termed the Histograms of Indexed Shapes (HIS) is used to localize a rectangular region that contains the ear [65], [54]. As previously

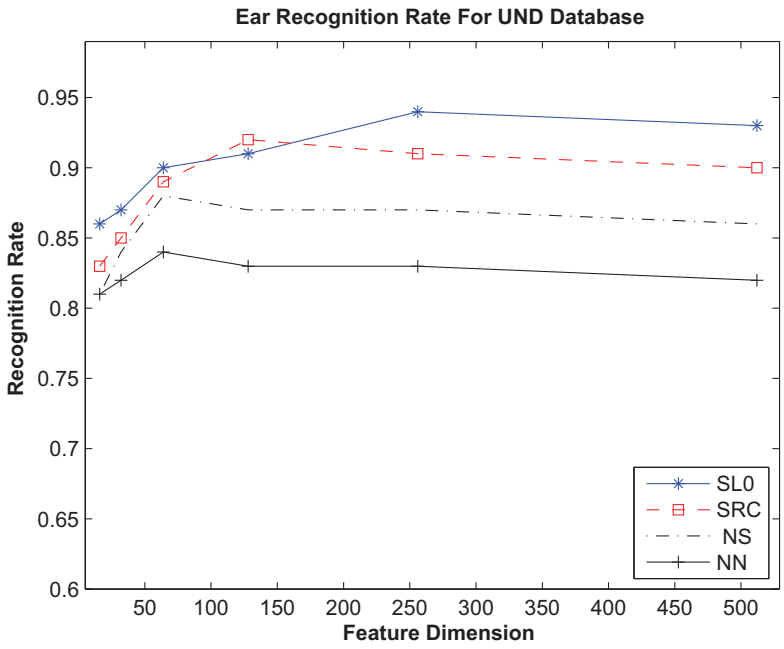


Figure 3.5: *Ear Recognition Rate on UND data set. A comparison of classifiers for different feature dimension*

stated, in sparse classification, the training samples are used to build the dictionary. In this database, some subjects only have a few images (e.g. 2 or 4 images), which is not suitable for sparse classification. Therefore, we selected 39 subjects that have more than 16 images . We used an equal number of images (10 images per subject) from each subject for training and the remaining images were used for testing. The results are shown in Fig. 3.5. The obtained results using proposed method (SL0) show that our method improves the overall recognition accuracy and just in one case (for 128 feature dimension) the recognition rate obtained by SRC is slightly better than results provided by our method.

In addition, The proposed method is compared to the state-of-the-art ear recognition algorithm employing sparse representation. In Table 3.1, the results based on

Table 3.1: Comparison Between l^1-Norm [3] and SL0

Method	TR4V2	TR3V3
l^1-Norm [3]	96.88%	91.67%
SL0	98.46%	93.66%

the SL0 algorithm are compared with the results of [3] that were obtained for the same database. Two protocols are proposed in [3] to evaluate the l^1-norm algorithm. The first protocol, termed TR4V2, uses 4 images for training and 2 images for testing. The second protocol, termed TR3V3, uses 3 images for training and 3 images for testing. In both protocols, the recognition rates of the SL0 algorithm are higher than those of the l^1-norm algorithm. It is obvious that the amount of training data is important since the testing data is represented by a linear combination of the training data. Hence, it is expected that the results will improve as the amount of training data increases.

3.5.3 Gender Classification Experiments

To evaluate our proposed method for the purpose of gender classification, two experiments were conducted on the UND and WVU data sets. In the experiments on WVU data set, the effect of different viewing angles on gender recognition is investigated. To our best knowledge, the effect of different viewing angles on ear classification was not studied before.

3.5.3.1 The First Experiment Using UND data set

The first gender classification experiment was performed on the UND data set. In this database, there are images for 241 male subjects and 174 female subjects.

Figure 3.6: *A few samples of extracted frames for one subject for different viewing angles*

As previously stated, in sparse classification, the training samples are used to build a dictionary, which is used during the recognition to represent a test sample as a linear combination of the training data samples. Since we are using majority voting for making a decision between the two categories, the number of training samples for males and females should be equal. However, in the UND data set, the number of images per subject are different. For example, for subject number one, who is a male, there are 22 images but for subject number eight who is a female, there are 12 images. We used all the images for each subject either for training or for testing. The reason is to avoid using test images for subjects that were used for training in order to make sure that the results are not biased. Therefore, in order to have equal number of training images for males and for females, the number of male and female subjects used for training were not the same. In this experiment, we have 60-40 split of the data set for training and testing.

Initially, the ear region is localized using our method described in [57]. Actually, a rectangular boundary around the ear region is extracted where the size of this boundary varies from subject to subject. Then ear regions are normalized such that all the rectangular boundaries have the size 140×90. In the feature extraction step, Gabor wavelets were extracted. Also, a uniform down-sampling by a constant value β was used to obtain a Gabor feature vector φ as shown by the red circles in Fig. 3.3 [1]. Gabor wavelets are extracted for 8 orientations and 5 spatial frequencies; one example of extracted features for one image is shown in Fig. 3.4. Given the size of the ear regions, 140×90 pixels, the Gabor wavelets coefficients are $140 \times 90 \times 8 \times 5 = 504000$. This feature vector is uniformly down sampled by a factor of 64 and a feature vector of size 7875 is obtained. Finally, using PCA, the number of features is reduced to 16, 32, 64, 128, 256 and 512. For each of these different feature sizes, the classification is performed using sparse representation.

In the proposed approach, for each test data, the sparsest coefficient vector $\widehat{\mathbf{x_0}}$ was obtained using the SL0 algorithm. Majority voting was then used to recognize the gender of the test subject. Fig. 3.7 shows the classification rates for the different feature dimensions: 16, 32, 64, 128, 256 and 512. In the same figure, results based on SL0 and majority voting are compared with Sparse Representation Classification (SRC), Adaboost, Nearest neighbour (NN) and Nearest Subspace (NS). For example, for feature size 512, the classification rates are 89.5, 88.5, 83.6, 64.1 and 59.2 for our method based on SL0, SRC, NN, Adaboost and NS, respectively. It is clear from the figure that the results based on sparse representation are far more robust than the results obtained from the other classifiers.

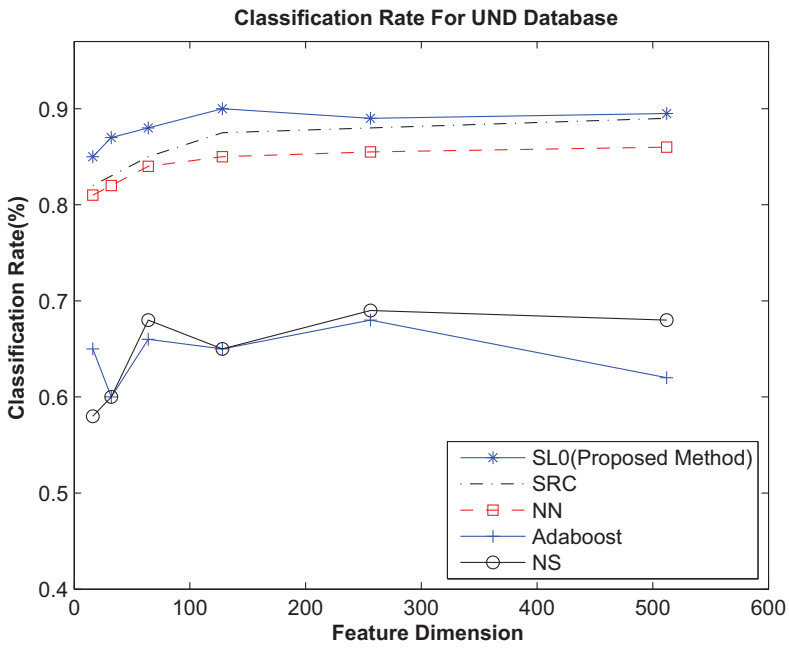


Figure 3.7: *classification Rate on UND data set. A comparison of classifiers for different feature dimension*

3.5.3.2 The Second Experiment Using WVU data set

The goal of this experiment was to examine the effect of the pose variations on the accuracy of the gender classification algorithm. For this purpose, we conducted two parts in this experiments. The WVU data set contains one video sequence for each subject. The video sequences start from the left profile of each subject (0 degrees) and terminate at the right profile (180 degrees) [68]. The length of each video sequence is about two minutes. A few subjects in the data set have eyeglass, earrings or part of the ear is occluded by hair. There are three subjects that have their ear fully occluded and these subjects were not used in the experiment. In the first part of the experiment, we perform an experiment using frames from different viewing angles for

training. However, in the second part, only zero degree frames are used for training and the rest of the data for testing.

In the first part of the experiment, 45 frames, which approximately cover the range of the camera positions from 0 to 44 degrees (*i.e.*, one frame for each degree), are extracted. Fig. 6.3 shows a few samples from a video sequence for one of the subjects. After extracting the frames, the ear region is detected and a bounding box around the ear is extracted. The ear detection is performed automatically based on our algorithm in [65]. Since the sizes of the extracted bounding boxes vary, we normalized the size to 120x80. The WVU database contains video clips for 402 subjects where 57 subjects are females and the rest are males. Since our algorithm requires an equal number of images in each class, all the female subjects (57 subjects) were used and 57 male subjects were chosen randomly. To build the dictionary, all the images for 34 male subjects and 34 female subjects were used. This means that all the images for one subject are used either for the training or testing and there is no overlap between the training and the test data. For each image, the feature vector is extracted using Gabor wavelets and is reduced in size to 16, 32, 64, 128, 256 and 512 using PCA. For instance, the size of the dictionary for feature vectors of size 32 is 32x3060 ($3060 = 45 \times 68$, where 68 is the number of subjects used for the training, 34 male and 34 female subjects, and 45 images were used for each subject). The classification rates for the different feature vector sizes are shown in Fig. 3.8. It is obvious that as the number of features increases, the classification rate increases as well. However, the complexity increases exponentially. In Figures 3.9 and 3.10 the classification rates for every test subject (male or female) are shown. For instance, in Fig. 3.9, all the images of the first test subject are correctly classified

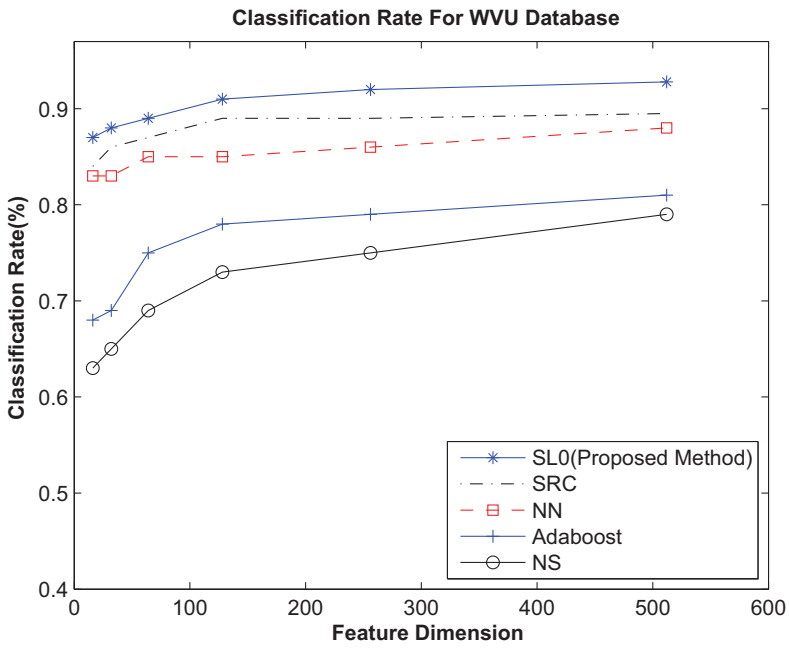


Figure 3.8: *classification Rate for WVU database*

as male. However, for subject 4, the classification rate is 91.11% which indicates 41 images out of total 45 images for this subject are classified as male and 4 images are wrongly classified as female. Furthermore, for the second test subject, in Fig. 3.10, the classification rate is close to zero which indicates that for this subject all the images are wrongly classified as male. In addition, to analyze the effect of the change in the viewing angle on the classification rate, the results of classification for different angles are observed. As mentioned, for each subject, there are 45 images associated to the different degrees from 0 to 44. In Fig. 3.11 the classification rates for the viewing angles from 0 degree to 44 degrees are shown. For example, the figure shows that for a viewing angle of 5 degrees, the classification rate is around 91%. As can be seen in Fig. 3.11, the classification rate is between 85% and 96%, which indicates that gender classification from different viewing angles is reliable.

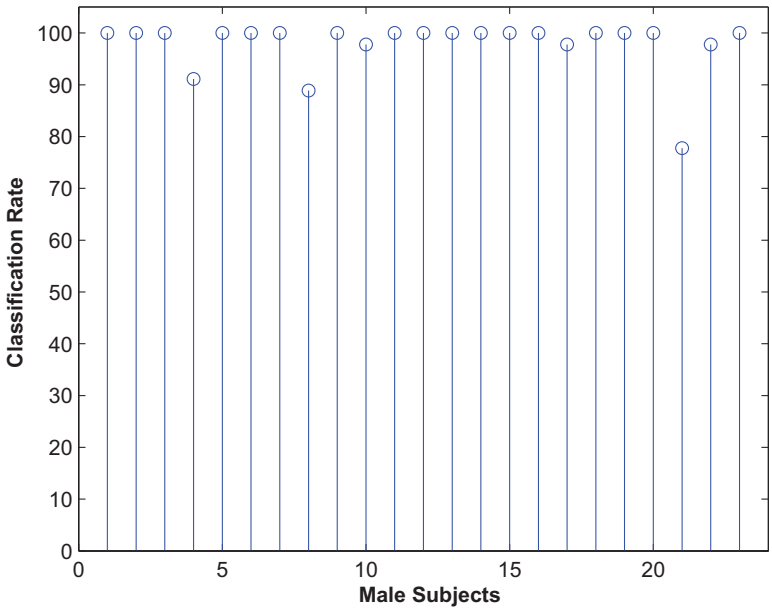


Figure 3.9: *Male Subjects classification Rate*

In the second part, the experiment is performed again on the WVU data set. Frame extraction and ear detection are the same as in the previous part. In this experiment, only extracted frames at zero degree are used for training and the extracted frames for subjects that were not used for the training are used for testing. Here, we use 34 female subjects and 34 male subjects (randomly chosen) and zero degree frames of the 68 subjects are used for training. On the other hand, the extracted frames from the rest of the subjects (from 0 degree to 44 degree) are used for testing. Fig. 3.12 and Fig. 3.13 show the classification rates for male and female subjects. The classification rate is 89.08% for male subjects and 84. 54% for female subjects. Actually, the recognition rate for female subjects is less as the ear part of female subjects is more occluded by hair or earrings which can affect the classification rate. In addition, in Fig. 3.14 the classification rates for different viewing angles are shown.

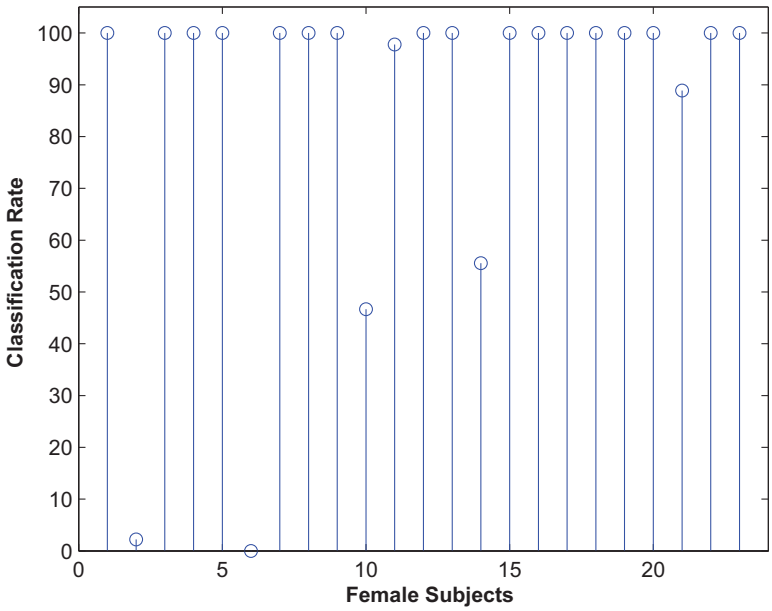


Figure 3.10: *Female Subjects classification Rate*

It is noticeable that as the difference in the viewing angle increases, the classification rate decreases. However, the classification rate is at least 81% which indicate that our proposed method has reliable gender classification performance from ear data for different viewing angles.

3.6 Conclusion

In this chapter, we presented a fully automated system for ear recognition and gender classification using sparse representation. The proposed method was evaluated on two data sets. The first experiment is performed on the UND collection J ear data set. Features were extracted using Gabor wavelets and a dictionary was constructed

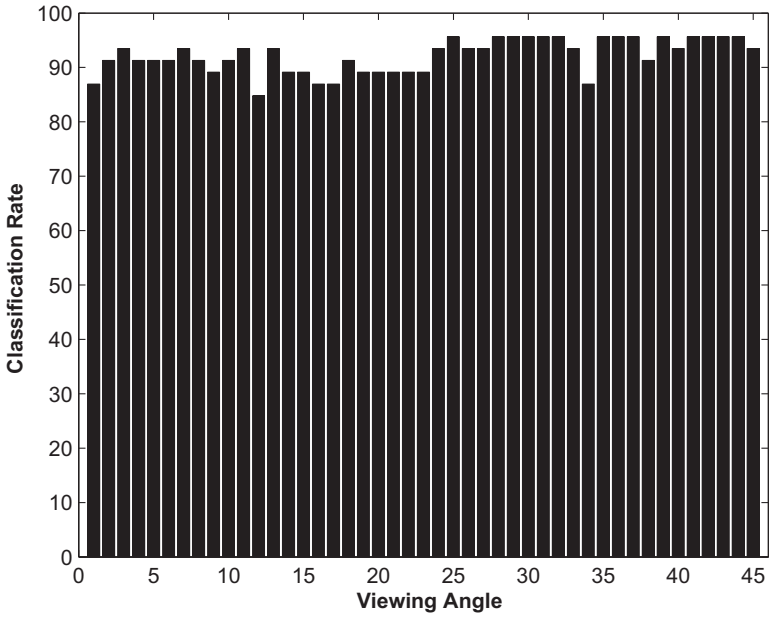


Figure 3.11: *Classification Rate For Different Viewing Angle*

using the extracted features from training subjects. The rest of the data set was used for testing. When 128 features were used, a classification rate of 89.49% was obtained using the SL0 algorithm for classification which is far better than other classifiers such as SRC, NN or NS. In the second experiment, the WVU data set, which contains a video sequence for each subject, was used. The video sequence starts from the left profile of the face and terminates at the right profile. In the first part of the experiment, 45 frames were extracted from 0 degree to 44 degrees for each subject and 60% of the subjects were used for training to build the dictionary. The rest of the subjects were used for testing and the best classification rate obtained by the proposed method is around 92%. In the second part of the experiment, only zero degree frames extracted from the training set were used for training. The test frames was extracted from video clips of other subjects for all the 45 degrees. These

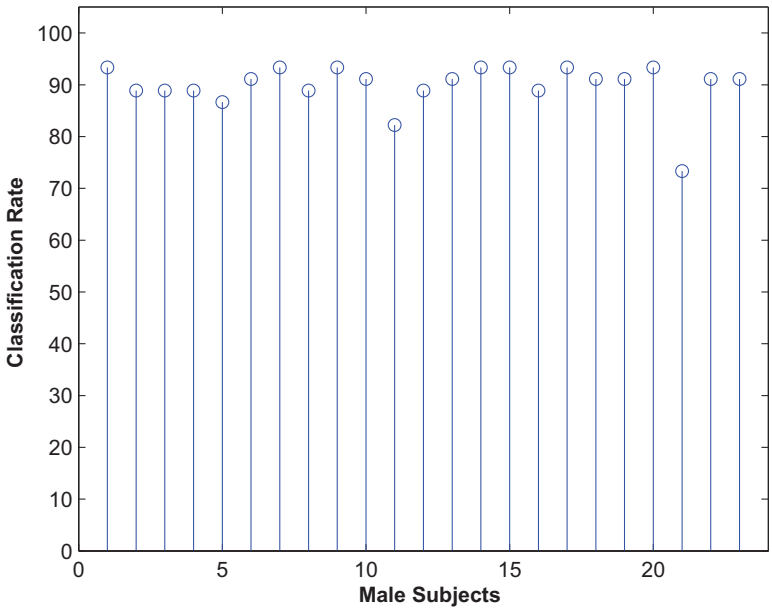


Figure 3.12: *Male Subjects Classification Rate (zero degree frames are used for training)*

results show that there is a graceful degradation of the results with the increase in the difference of the viewing angle. The obtained results using the proposed method were far more robust than the results obtained from the other classifiers that were used for comparison. In the future, we plan to fuse facial and ear features for the purpose of gender classification.

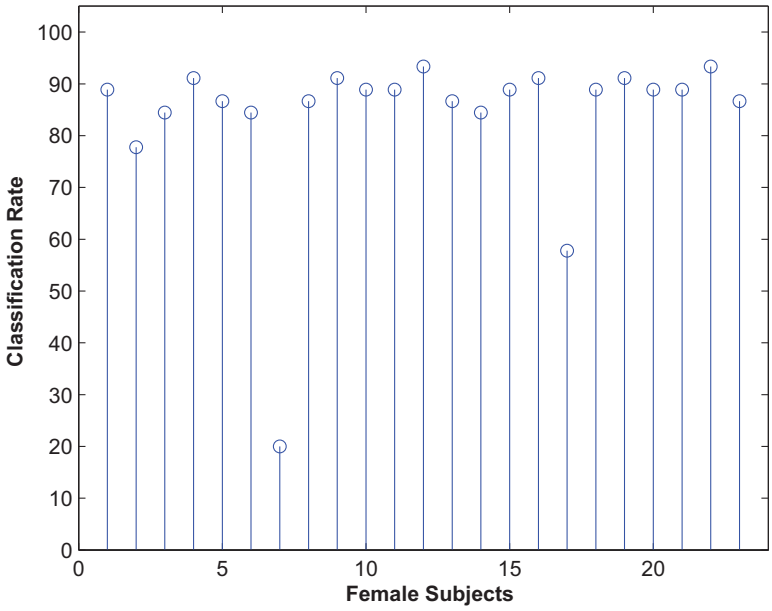


Figure 3.13: *Female Subjects Classification Rate (zero degree frames are used for training)*

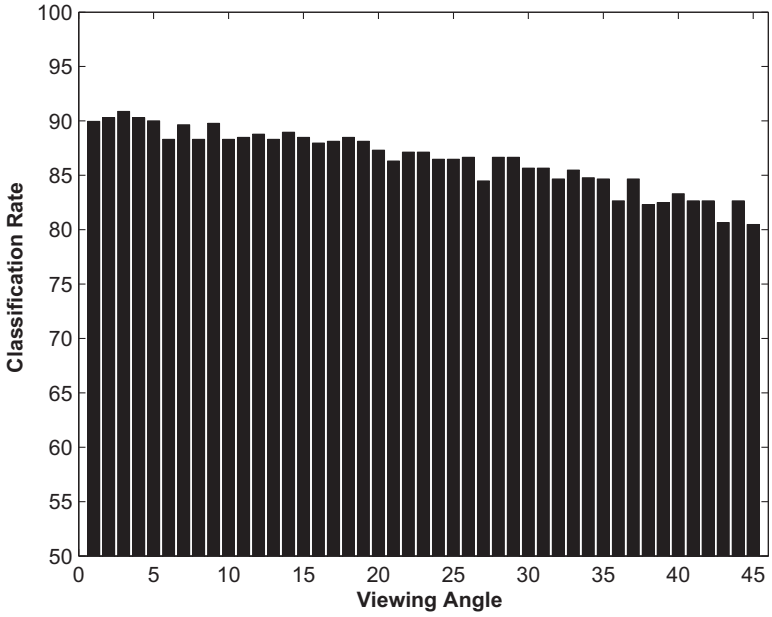


Figure 3.14: *Classification Rate For Different Viewing Angle (zero degree frames are used for training)*

CHAPTER 4

Gender Classification Using Facial Images and Basis Pursuit

4.1 Overview

Gender classification is an important task in social activities and communications. In fact, automatically identifying gender is useful for many applications, *e.g.* security surveillance [47] and statistics about customers in places such as movie theaters, building entrances and restaurants [48]. Automatic gender classification is performed based on facial features [49], voice [51], body movement or gait [52].

Most of the published work in gender classification is based on facial images. Moghaddam *et al.* [59] used Support Vector Machines (SVMs) for gender classification from facial images. They used low resolution thumbnail face images (21×12 pixels). Wu *et al.* [60] presented a real time gender classification system using a Look-Up-Table Adaboost algorithm. They extracted demographic information from human faces. Gollomb *et al.* [49] developed a neural network based gender identification system. They used face images with resolution of 30x30 pixels from 45 males and 45 females to train a fully connected two-layer neural network, SEXNET. Cottrell and Metcalfe [61] used neural networks for face emotion and gender classification from facial images.

Gutta and Wechsler [62] used hybrid classifiers for gender identification from facial images. The authors proposed a hybrid approach that consists of an ensemble of RBF neural networks and inductive decision trees. Yu *et al.* [52] presented a study of gender classification based on human gait. They used model-based gait features such as height, frequency and angle between the thighs. Face-based gender classification is still an atractive research area and there is room for developing novel algorithms that are more roubst, more accurate and fast.

In this chapter, we present a gender classification system based on 2-D facial images and sparse representation. This chapter is organized as follows: In Section II, we present a brief mathematical explanation of the sparse representation concept and the proposed method based on basis pursuit to obtain the sparsest solution. Section III presents experimental results that demonstrate the performance of the proposed method in terms of recognition. Conclusions and future research directions are discussed in Section IV.

4.2 Classification based on Sparse Representation

Underdetermined systems appear in different important areas such as signal processing, statistics, pattern recognition and image processing. Sparse representation is a relatively new approach to solve underdetermind systems. In this section, we briefly explain the concept of sparse representation based on Basis Pursuit. The proposed approach for finding the sparsest solution based on basis pursuit is described, and Gabor wavelets, which we used for extracting the feature vectors, are discussed.

4.2.1 Sparse Solution Based on Basis Pursuit

Basis pursuit was introduced in the 1970s, and then studied mathematically in the 1990s by Chen and Donoho [69]. To solve the underdetermined system of equations $\mathbf{y} = \mathbf{Ax}$, l^2 minimization is easy to compute, but not useful in recognition. In fact, for recognition purposes, minimizing the l^0 norm provides the best solution since the test data is related to only one of the subjects in the training set. In fact, most of the components of $\mathbf{x}$ should be zero or close to zero. However, the l^0 norm is not a continous function. Since l^0 norm minimization is not a convex optimization problem, it is not easy to obtaine the solution. On the other hand, we can use l^1 norm minization, which is convex, to find $\mathbf{x}$. In the l^1 norm minimization, a cost is assigned to each atom that we use in our representation. Actually, there is no charge for the norm when it gives a zero coefficient. The BP finds the best solution of $\mathbf{x}$ by minimizing the l^1 norm of the $\mathbf{x}$ as follows:

$$\widehat{\mathbf{x}}_1 = argmin \quad \|\mathbf{x}\|_1 \quad subject\ to \quad \mathbf{y} = \mathbf{Ax} \tag{4.1}$$

Where $\|\mathbf{x}\|_1$ is the l^1 norm. To find $\widehat{\mathbf{x}}_1$, since the nonzero coefficients correspond to columns of the dictionary, it is possible to use the indices of the nonzero components of $\widehat{\mathbf{x}}_1$ to identify the columns of A that are necessary to represent the test image. l^1 norm assigns a cost to each atom that is used in representation. For example, the norm will not be penalized when it gives a zero coefficient, but it should be charged proportionally for small and large coefficients.

Since $\|\mathbf{x}\|_1 = |x_1| + ... + |x_n|$ and we can rewrite equation 7.2 as:

$$minimize \|\mathbf{x}\|_1 = |x_1| + ... + |x_n| \quad subject\ to \quad \mathbf{y} = \mathbf{Ax} \tag{4.2}$$

Because $|x_1| + ... + |x_n|$ is a nonlinear function, the optimization problem can not be solved using linear programming methods. To make this function linear, nonlinearities should change to constraints by adding new variables as follow:

$$minimize \quad t_1 + t_2 + ... + t_n \quad subject\ to \tag{4.3}$$

$$|x_1| \leq t_1, \quad .... \quad , |x_n| \leq t_n \quad and \quad \mathbf{y} = \mathbf{Ax}$$

Where, $t_1, ..., t_n$ are non-negative constants. In this formulation, the objective function is linear and it is possible to solve this problem using linear programming.

4.2.2 Using Sparse Representation for Classification

For a test data, $\mathbf{y}$, belonging to the i^{th} class, it is assumed that the non-zero elements of $\widehat{\mathbf{x_1}}$ will correspond to the training data samples from the i^{th} class. However, due to noise and representation errors, there will be extraneous non-zero elements corresponding to training samples from other classes.

In [20], they presented an approach for the decision making step based upon the obtained $\widehat{\mathbf{x}}_1$ by computing the error between $\mathbf{y}$, the original data, and $\widehat{\mathbf{y}}_\mathbf{i}$, the approximation obtained through the sparse representation. For each class i and $\mathbf{x} \in \mathbf{R^n}$, vector $\delta_i(\mathbf{x}) \in \mathbf{R^n}$ represents the coefficients that are associated with class i. Using this definition, approximated test data $\widehat{\mathbf{y}}_\mathbf{i}$ is given as:

$$\widehat{\mathbf{y}}_\mathbf{i} = A\delta_i(\widehat{\mathbf{x_1}}) \tag{4.4}$$

Recognition was performed by assigning the test data to the class that minimizes the residual between $\mathbf{y}$ and $\widehat{\mathbf{y}}_\mathbf{i}$ as follows:

$$\underbrace{min}_{i} \quad r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\delta_\mathbf{i}(\widehat{\mathbf{x}}_\mathbf{1})\|_\mathbf{2} \tag{4.5}$$

where $r_i(\mathbf{y})$ is the residual distance for class i. This signifies that the classification is performed based on the best approximation and least error [20].

Here, we propose a new approach to perform classification using $\widehat{\mathbf{x}}_1$. In gender classification, there are only two classes and the dictionary contains training face images for males and females as representatives of these two classes. The obtained elements of $\widehat{\mathbf{x}}_1$ are the coefficients associated with each training face image and we can divide $\widehat{\mathbf{x}}_1$ into two vectors, $\mathbf{x_1}$ and $\mathbf{x_2}$, where $\mathbf{x_1}$ contains the coefficients associated with males and $\mathbf{x_2}$ contains the coefficients associated with females. $\widehat{\mathbf{x}}_1$ can be written as follows:

$$\widehat{\mathbf{x}}_1 = \begin{bmatrix} \mathbf{x_1} \\ \\ \mathbf{x_2} \end{bmatrix}$$

The length of the $\widehat{\mathbf{x}}_1$ is m and the number of training samples for males and females are equal. Hence, the length of $\mathbf{x_1}$ and the length of $\mathbf{x_2}$ is $m/2$.

Let x_{max} be the maximum value of the $\widehat{\mathbf{x}}_1$ elements ($x_{max} = max(\widehat{\mathbf{x}}_1)$). Then, a threshold x_{max}/τ, where ($\tau \geq 1$) is defined. The elements in $\mathbf{x_1}$ and $\mathbf{x_2}$ whose values are more than the threshold are counted. The classification is performed based on the majority vote of the coefficients.

4.3 Gabor Wavelets

The Gabor filters (kernels) with orientation μ and scale ν are defined as [1]

$$\psi_{\mu,\nu} = \frac{\| k_{\mu,\nu} \|^2}{\sigma^2} e^{\left(\frac{-\|k_{\mu,\nu}\|^2 \|z\|^2}{2\sigma^2}\right)} \left[e^{ik_{\mu,\nu}z} - e^{-\sigma^2/2} \right] \tag{4.6}$$

where $z = (x, y)$ is the pixel position, and the wave vector $k_{\mu,\nu}$ is defined as $k_{\mu,\nu} = k_\nu e^{i\phi_\mu}$ with $k_\nu = k_{max}/f^\nu$ and $\Phi_\mu = \pi\mu/8$. k_{max} is the maximum frequency,

and f is the spacing factor between kernels in the frequency domain. The ratio of the Gaussian window width to wavelength is determined by σ. Considering Eq. 4.6, the Gabor kernels can be generated from one wavelet, *i.e.,* the mother wavelet, by scaling and rotation via the wave vector $k_{\mu,\nu}$ [67]. In this work, we used five scales, $\nu \in \{0, ..., 4\}$ and eight orientations $\mu \in \{0, ..., 7\}$. We also used $\sigma = 2\pi$, $k_{max} = \pi/2$ and $f = \sqrt{2}$.

4.4 Experiments and Results

The FERET database [70] is used to validate the proposed method. Images are frontal faces at a resolution of 256x384 with 256 gray levels. All the images are pre-processed before applying the algorithm. First, the automatic eye-detection method is applied based on the [71] and the distance, d, between the 2 eye corners is measured. Then, the middle point between the 2 eye corners is found and the image is cropped by the size of $2d \times 2d$. Then all images are resized to 128x128. A few sample face images for both male and female subjects are shown in Fig. 4.1. In this database, there are 250 male subjects and 250 female subjects. As previously stated, in sparse classification, the training samples are used to build a dictionary, which is used during the classification to represent a test sample as a linear combination of the training samples. Since we are using majority voting for making a decision between the two categories, the number of training samples for males and females should be equal. In addition, to compare our results with other methods, especially [72], four experiments are conducted with different number of subjects used for training in each experiment, sizes of: 50, 100, 150 and 200 subjects were used for training. In each

Figure 4.1: *Sample images for both males and females in FERET database*

experiment, the remaining subjects are used for testing. For instance, when using 200 subjects for training (100 male subjects and 100 female subjects), the other 300 subjects are used for testing.

In the feature extraction step, Gabor wavelets are extracted for 8 orientations and 5 spatial frequencies. Finally, using PCA, the number of features used to represent each image is reduced to 128. In the proposed approach, for each test data, the sparsest coefficient vector $\widehat{\mathbf{x_1}}$ is obtained based on the basis pursuit. Majority voting is then used to recognize the gender of the test subject. We provide a comparison of the experimental results with other gender classification systems applied to the same dataset. Table. 4.1 shows the classification rates for 4 different training set sizes. The results of our proposed method (PCA + BP) are compared with the results of the methods proposed by Jain *et al.* [72], in which the authors evaluated their method on the FERET database. They used Independent Component Analysis (ICA) to represent each image as feature vector in a low dimensional subspace. In addition, they used different classifiers such as cosine classifier (COS), linear discriminant classifier (LDA) and the support vector machine (SVM). The best result reported in [72]

Table 4.1: Performance comparison to other gender classification systems based on facial images

Training Set Size	COS	LDA	SVM	SRC	PCA+BP (Proposed Method)
50	60.67%	64.67%	68.30%	68.88%	68.88%
100	71.67%	73.67%	76.00%	76.00%	76.25%
150	80.33%	83.00%	86.67%	86.85%	88.57%
200	85.33%	93.33%	95.67%	96.33%	97.66%

is 95.67% accuracy using SVM with ICA. Furthermore, the results for conventional sparse representation based classification (SRC) [20] are reported in Table. 4.1 which show our modification was helpful in gender recognition. The experimental results in this chapter indicate that our proposed method using sparse representation and PCA obtained higher performance of correct classification rate on the same data set. To our best knowledge, better results for gender classification on FERET database is not reported since 2005. Moreover, in [73], authors used 661 images from FERET database, for 248 subjects. The best obtained result for gender classification in that paper is 90% for feature dimension 11,520. However, we obtained a classification rate of 97% for 512 feature dimension and for 500 subjects.

4.5 Conclusion

In this chapter, we presented a method for gender classification, from facial images, using sparse representation. Basis pursuit method was used to formulate the problem in order to find the sparsest solution. The experiments were conducted on the FERET data set containing 500 subjects (250 male and 250 female subjects). Fea-

tures were extracted using Gabor wavelets, and a dictionary was constructed based on the extracted features from a training set. The rest of the data set was used for testing. We compared the proposed method with previous methods that used the same data set, performance of our the presented method is better than the previous reported methods.

CHAPTER 5

Classification based on Weighted Sparse Representation using Smoothed l^0 Norm with Non-negative Coefficients

5.1 Overview

Recently, the sparse representation based classification (SRC) has been successfully used in face recognition. In practical applications, robust face recognition is a challenging task due to the significant variations that can be encountered in face images. Wright et. al. [20] proposed a face recognition algorithm, based on sparse representation and l^1 norm minimization, which is robust towards variations in lighting conditions, facial expressions and partial occlusions. In fact, the sparse non-zero coefficients should concentrate on the training samples with the same class label as the query sample. On the other hand, due to the rich information contained in ear images, the ear is becoming an important biometrics for recognition and identification. In this chapter, we use the smoothed l^0 norm algorithm with non-negative constraints on the coefficient vector. In addition, we obtain weights, using mutual information, for each training sample. Therefore, the atoms in the dictionary are not treated uniformly and the use of weights helps to narrow the search space for the coefficient

vector. Actually, these weights help in reducing the chances of the algorithm to be trapped in local extremums because it assigns a small weight to irrelevant subjects. The proposed method for classification and recognition, Weighted Sparse Representation using Smoothed l^0 Norm with Non-negative Coefficients, is described in the next section. To the best of our knowledge, our proposed method is the first to use mutual information for obtaining weights and to use SL0 with weights and non-negative constraints. To evaluate the proposed method, several experiments are conducted on face and ear biometrics. The obtained results for face and ear recognition using standard databases show that the proposed algorithm has accurate and robust performance.

5.2 The Proposed Method

In this section, we provide a brief overview of Mutual Information and we detail our proposed method, Classification based on Weighted Sparse Representation using Smoothed l^0 Norm with Non-negative Coefficients.

5.2.1 Mutual Information

In the proposed method, the mutual information between the query sample and each of the training samples is calculated and the obtained results are used as weights of the atoms in our proposed method. In fact, the query sample is related to one of the training classes and the mutual information between the query sample and the training samples from the same class are expected to be high. The issue that we are trying to deal with is that the sparse representation may reconstruct the query sample using training images which are not from the class of the query sample and as

a consequence might lead to wrong classification. Adding useful information between the query sample and each of the training samples can reduce the search space and speed up the convergence to the optimal solution.

In information theory, mutual information (MI) can be applied for evaluating any arbitrary dependence between random variables such as signals and images. Actually, the MI between two random variables $\mathbf{A}$ and $\mathbf{B}$ is a measure of the amount of information between them. For example, if $\mathbf{A}$ and $\mathbf{B}$ are independent, the MI will be close to zero, whereas if both variables are closely related, the MI value will be large. The MI of two random variables $\mathbf{A}$ and $\mathbf{B}$ is defined as:

$$MI(\mathbf{A},\mathbf{B}) = H(\mathbf{A}) - H(\mathbf{A}|\mathbf{B}) = H(\mathbf{A}) + H(\mathbf{B}) - H(\mathbf{A},\mathbf{B}) \tag{5.1}$$

where $H(\mathbf{A})$ and $H(\mathbf{B})$ are the entropies for variables $\mathbf{A}$ and $\mathbf{B}$, respectively. $H(\mathbf{A}|\mathbf{B})$ is the conditional entropy and $H(\mathbf{A},\mathbf{B})$ is the joint entropy of $\mathbf{A}$ and $\mathbf{B}$. The MI formulation based on the probability functions is:

$$MI(\mathbf{A},\mathbf{B}) = E\{log\frac{p_{\mathbf{A},\mathbf{B}}(\mathbf{A},\mathbf{B})}{p_{\mathbf{A}}(\mathbf{A})p_{\mathbf{B}}(\mathbf{B})}\}$$

$$= \int\int p_{\mathbf{A},\mathbf{B}}(\mathbf{A},\mathbf{B})log\frac{p_{\mathbf{A},\mathbf{B}}(\mathbf{A},\mathbf{B})}{p_{\mathbf{A}}(\mathbf{A})p_{\mathbf{B}}(\mathbf{B})} \tag{5.2}$$

where $E\{.\}$ is the expectation, $p_{\mathbf{A}}(\mathbf{A})$ and $p_{\mathbf{B}}(\mathbf{B})$ are the marginal probability distributions, and $p_{\mathbf{A},\mathbf{B}}(\mathbf{A},\mathbf{B})$ is the joint probability distribution.

We define the diagonal weight matrix $\mathbf{W}^i$, which contains the weights between the query and the training samples of subject i, as follows:

$$w^i_{j,j} = 1/MI(\mathbf{I_y}, \mathbf{I_x}^i_j) \tag{5.3}$$

$$\mathbf{W^i} = \begin{bmatrix} w^1_{1,1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & w^i_{n_i,n_i} \end{bmatrix} \tag{5.4}$$

where $w^i_{j,j}$ is the j^{th} diagonal element of the matrix $\mathbf{W}^i$, I_y is the query image and the $\mathbf{I}_{\mathbf{x}}{}^i_j$ is the j^{th} training image from subject i. The weight matrix W, for all the training samples, is defined as follow:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W^1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{W^k} \end{bmatrix} \tag{5.5}$$

where k is the number of subjects.

5.2.2 The Proposed Algorithm

In this section, the proposed method, weighted sparse representation using SL0 with non-negative coefficients, is described. The following l^0 norm minimization problem is considered:

$$(NW \quad l^0): \quad \widehat{\mathbf{x}}_0 = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{Wx}\|_0$$

$$\textit{Subject to} \quad \mathbf{y} = \mathbf{Axand} \quad \mathbf{x} \geq \mathbf{0} \tag{5.6}$$

where W is the weight matrix obtained using mutual information. As mentioned before, the obtained coefficients can be used to reconstruct the query from the training

samples. With the constraint that the coefficients, i.e., components of X, are either 0 or positive, the coefficients represent the contribution of the individual samples in constructing the query. Without this constraint on coefficients, the query sample can be reconstructed by adding and subtracting contributions from the training samples. This is contradictory to the intuitive notion of combining parts to form a whole [74], [19]. In fact, using non-negative coefficients can provide insight into the similarity between the query sample and each of the training samples. Thus, the samples associated with the nonzero coefficients should have more or less similar features of the test sample.

On the other hand, in sparse representation, the query sample may be reconstructed by training samples which are not in the same class and thus produce not accurate classification results. We use weights in our proposed method to improve sparse representation results. Information about the similarity between the query sample and each training sample can be useful into finding more discriminative and accurate coefficient vector. Hence, using accurate coefficient vector, the reconstruction of query sample will be more precise and the recognition will be improved. In fact, if the similarity between the query sample and a training sample is low which means the query sample and the training sample probably are not related to the same subject (the weight will be high, $w = 1/MI(I_X, I_Y)$), the corresponding coefficient will be small (this training sample will not be effective in reconstruction). Basically, in SRC, all the atoms in the dictionary are treated uniformly and larger coefficients will penalized more than smaller one in optimization process. Hence, the weights are adopted to counteract the influence of the magnitude of coefficients on the penalty

function. In this method, training samples which are more similar to query sample are distinguished using mutual information.

5.3 Feature Extraction and Dimensionality Reduction

One of the transform based methods for feature extraction is the Gabor Wavelets which is extensively used in many applications of computer vision, including biometrics. A 2-D Gabor wavelet representation is presented in [75] for the classification of facial images. In this section, we provide a brief explanation of Gabor wavelets and its formulation. Furthermore, we use Histogram of Oriented Gradients (HOG) descriptor for ear recognition, which was first proposed and efficiently used for object detection and image retrieval [76], especially when illumination variations are present. Actually, it is considered as one of the best features for the dense encoding of 2D image regions, and has been successfully used in pedestrian detection and object classification tasks [77].

5.3.1 Histogram of Oriented Gradients

We use the HOG feature for ear recognition. It was demonstrated that this feature achieves excellent performance in image retrieval [76] and 2D object detection tasks [77]. The HOG feature descriptor is in fact a dense version of the SIFT feature

- **Dictionary:** Constructing the over complete dictionary ($\mathbf{A}$) using extracted features from the training samples.
- **Weights:** Obtain the weights matrix, $\mathbf{W}$, using mutual information between the query sample and each of the training samples.
- **Initialization:**

  Obtain $\mathbf{x'_0}$, the solution that minimizes the l^2 norm of $\mathbf{y} = \mathbf{Ax}$ using pseudo-inverse of $\mathbf{A}$ ($\mathbf{A}^\dagger$), $\mathbf{x'_0} = \mathbf{A^T}(\mathbf{AA^T})^{-1}$.

  Select a decreasing sequence for σ, $[\sigma_1 \dots \sigma_K]$.
- **Iteration:** For $k = 1, \dots, K$:
  1. Let $\sigma = \sigma_k$.
  2. Find $\mathbf{x}$ by maximizing the function $\mathbf{F}_\sigma(\mathbf{x}) = \sum_i \exp(-x_i^2/2\sigma^2)$ on the feasible set $\mathcal{S} = \{\mathbf{x}|\mathbf{y} = \mathbf{Ax}\}$ using L iterations of the steepest ascent algorithm:
     - Initialization: $\mathbf{x} = \mathbf{x'}_{k-1}$.
     - For $\ell = 1 \dots L$ (loop L times):
       (a) Let, $\bigtriangledown_{\mathbf{x}}\mathbf{F}_\sigma(\mathbf{x}) \propto -[x_1 exp(-x_1^2/2\sigma^2) \quad ... \quad x_n exp(-x_n^2/2\sigma^2)]^T$
       (b) Let $\mathbf{x} \leftarrow \mathbf{x} - \eta\mathbf{W} \times \bigtriangledown_{\mathbf{x}}\mathbf{F}_\sigma(\mathbf{x}) - \lambda\mathbf{x}$ (η and λ are small positive constants).
       (c) Project $\mathbf{x}$ back onto the feasible set:
       $$\mathbf{x} \leftarrow \mathbf{x} - \mathbf{A}^\dagger(\mathbf{Ax} - \mathbf{y})$$
  3. Set $\mathbf{x'}_k = \mathbf{x}$.
- **$\mathbf{l^0}$ Norm Solution:**, the sparsest solution is $\widehat{\mathbf{x}}_0 = \mathbf{x'}_K$.
- **Recognition:** Finding the smallest residual
$$\underbrace{min}_{i} \quad r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\delta_\mathbf{i}(\widehat{\mathbf{x}}_\mathbf{0})\|_\mathbf{2}$$

Figure 5.1: Our Proposed Method

descriptor, i.e. SIFT descriptor computed on a dense grid. The HOG has several advantages as it captures edge or gradient structure which is very characteristic of local shape. In fact, it does so in a local representation with an easily controllable degree of invariance to local geometric and photometric transformations [77]. First, HOG decomposes an image into small rectangular cells. Then, it computes the histogram of oriented gradients in each cell. Generally, to build a feature vector using HOG, the cell histogram of each pixel within the cell casts a weighted vote, according to the gradient l^2 norm, for an orientation-based histogram channel. The gradient strengths were locally normalized over each cell, in order to account for changes in illumination and contrast [78]. There are two main parameters to compute HOG, the number of rectangular cells and the number of histogram bins per cell. For example, 10 bins and 9 rectangular cells were then concatenated to make a 90-dimensional feature vector.

5.4 Experiments

In this section, several experiments are performed to demonstrate the effectiveness of the proposed method in terms of recognition accuracy. We present our experimental results on two publicly available databases for face and ear recognition. We evaluate our method using several feature extraction methods and compare it with other classification algorithms such as Nearest Neighbour (NN) and Nearest Subspace (NS), as well as the original SRC and SL0.

5.4.1 Face Recognition

We selected the Extended Yale B database [79] which contains face images captured under large illumination variations. This database is commonly used to evaluate the performance of illumination invariant face recognition methods. The database consists of 64 face images per subject for 38 subjects under different illuminations. Frontal pose images were selected for our experiment. To simulate the results from [20], images were cropped and normalized to the size of 192×168 and half of the images were randomly selected for training and the rest for testing. The experiments were performed with feature space dimensions of 30, 56, 120 and 504 (corresponded to downsampling the images with ratios of $1/32, 1/24, 1/16$ and $1/8$). Example images of one person in frontal pose are shown in Fig. 5.2. The face recognition results without any preprocessing are illustrated in Table 7.1. The proposed method achieves the best recognition rate compared to other methods when using 56, 120 and 504 features. Only, when the number of features is 30, the recognition rate of the NS approach is slightly better than the proposed method. However, the recognition results of the proposed method when using more features is much better than NS.

Furthermore, in order to evaluate the effect of the number of training samples on the recognition rate, the number of the features is fixed at 56 and the number of training images per subject is decreased. We used 30, 25, 20, 15 and 10 images per subject for training and the rest of the images for testing. The obtained results are shown in Table 7.2. In this table, the recognition rate of our method is better than

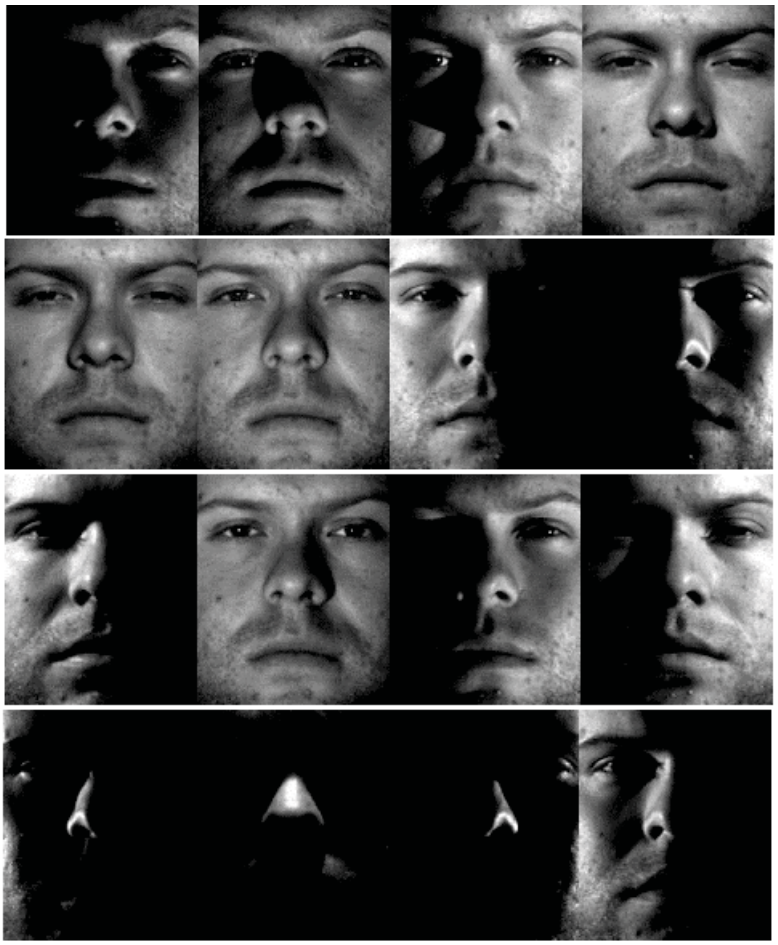


Figure 5.2: *Sample Frontal Face Images for One Subject in Extended Yale B Database*

all the other methods, except for NS when the number of training samples is 10; in this case the NS has a slightly better performance than our method.

We carried a third experiment to demonstrate the robustness of the proposed method with fewer number of training images per subject and fewer number of features, the number of the training samples was decreased from 32 to 25 and 20 images which correspond to 40% and 30% of the total number of images per subject, respectively, and the rest of the images were used for testing and the number of features

Table 5.1: Face Recognition Rates on Extended Yale B database (50% for training)

Feature Dimension	30	56	120	504
NN	69.3%	72.8%	78.5%	79.5%
NS	79.6%	84.1%	88.7%	90.8%
SRC	75.7%	84.8%	93.9%	96.8%
SL0	74.1%	82.6%	93.3%	95.5%
The proposed Method	78.5%	86.7%	95.3%	97.9%

were 32, 56, 120 and 512. The obtained results as illustrated in Fig. 5.3 and Fig. 5.4 for different number of features, specially for low number of features, show that the proposed method is more robust when the number of training samples are reduced.

5.4.2 Ear Recognition

The University of Notre Dame (UND) dataset collection J2 is used to test the proposed method for 2D ear recognition. The Histograms of Indexed Shapes (HIS), a shape-based feature set, is used to localize a rectangular region that contains the ear [65], [54]. A few profile face images from the UND dataset are shown in Fig. 5.5. As previously stated, in sparse classification, the training samples are used to build the dictionary. In this dataset, some subjects only have a few images (e.g. 2 or 4

Table 5.2: Face Recognition Rates on Extended Yale B database (The number of features is 56)

The Number of Training Images Per Subject	30	25	20	15	10
NN	64.9%	60.0%	56.9%	52.1%	42.7 %
NS	84.7%	84.8%	83.1%	80.2%	73.0 %
SRC	84.3%	82.6%	81.0%	78.5%	69.8 %
SL0	81.3%	79.7%	77.6%	74.3%	66.6%
The Proposed Method	85.7%	85.5%	84.5%	81.1%	72.5 %

images), which is not suitable for sparse representation classification. Therefore, we selected 39 subjects that have more than 16 images each. We used an equal number of images (10 images per subject) from each subject for training and the remaining images were used for testing. Gabor Wavelets were used for feature extraction and the number of features were reduced using PCA. As mentioned previously, the equation $\mathbf{y} = \mathbf{Ax}$ should be under determined and the number of columns in the dictionary should be more than the number of rows. Since, we used 10 images per subject for training, the number of columns in the dictionary is 390 and the number of features should be less than this number. Using PCA we reduced the number of features to 16, 32, 64, 128 and 256. Fig. 6.2 shows a comparison of the recognition results of our algorithm, SRC and SL0 when using different number of features. The obtained results using our proposed method show significant improvement in the recognition accuracy.

To evaluate the robustness of the proposed method for ear recognition, the number of training samples is decreased from 10 samples to 5 samples. Since there are 39

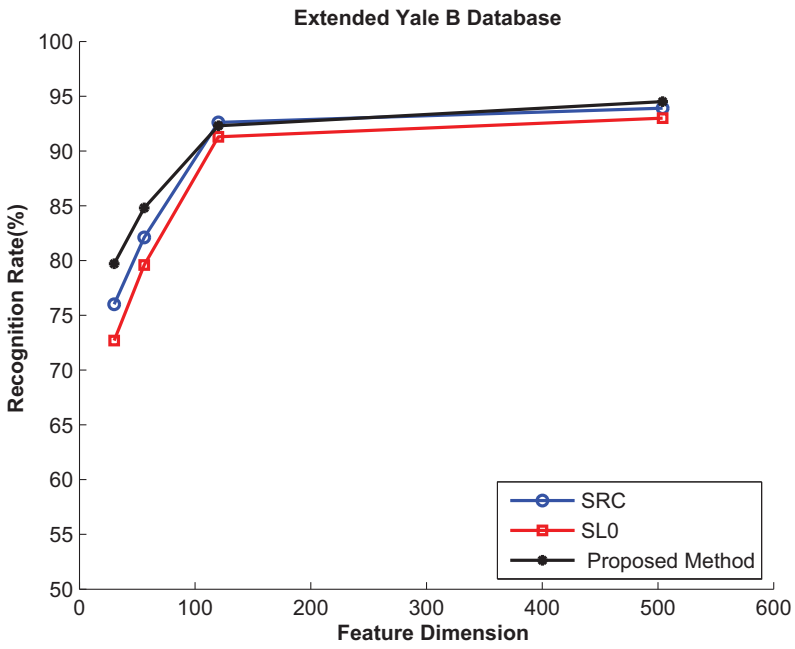


Figure 5.3: *Face Recognition Rates on Extended Yale B database. 25 images per subject are used for training and the rest for testing.*

subjects and 5 sample images per subject are used for training, there are 195 atoms (columns) in the dictionary. Therefore we obtained results for 16, 32, 64 and 128 features, as shown in Fig. 5.7. The results obtained using our method are more robust than SRC and SL0. For example, for lower dimension feature vectors, *e.g.*, 16 features, the results from our method are far better than the results for SL0 and are slightly better than those of SRC. However, for higher number of features, *e.g.*, 128 features, the SRC results are not as good as SL0 and our method. In order to check the performance of our algorithm with other types of features, we used histogram of oriented gradients (HOG) for feature extraction. The number of rectangular cells are fixed to 9 (3×3) and the number of histogram bins per cell are 4, 6, 8, 10, 12, 14 and 16. The extracted feature vectors corresponding to the different numbers of histogram

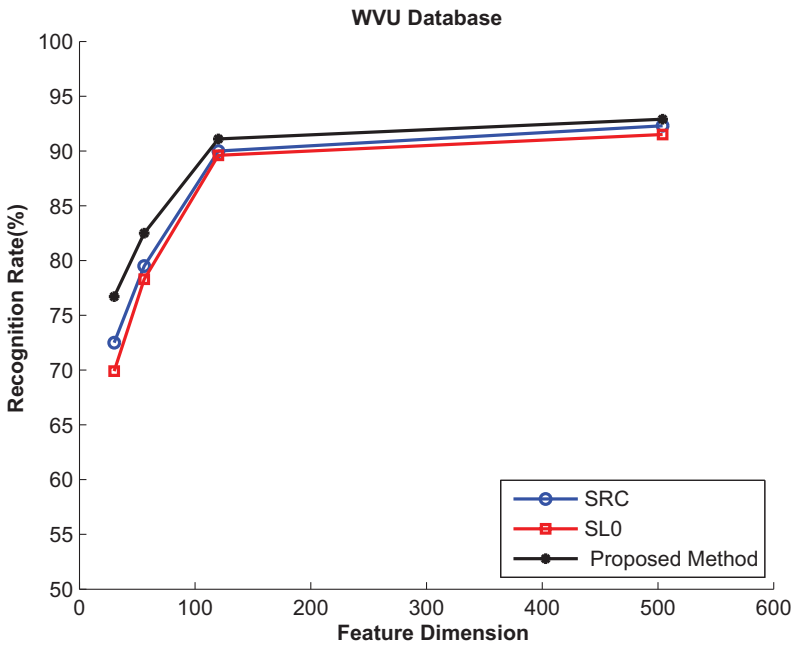


Figure 5.4: *Face Recognition Rates on Extended Yale B database. 20 images per subject are used for training and the rest for testing.*

bins per cell are 36, 54, 72, 90, 108, 126 and 144. The recognition rates for SRC, SL0 and our method are shown in Fig. 5.8 for different numbers of features. These results still show that the performance of the proposed method is better thank SRC and SL0 for HOG features.

5.5 Discussion

In this chapter, three recognition methods based on sparse representation: SRC, SL0 and our proposed method, were compared. SRC uses the l^1 norm optimization solver (l^1 solver) and the two other methods use the smoothed l^0 norm optimization solver (l^0 solver). Based on the experiments and results obtained from the Extended

Figure 5.5: *Profile Image Samples From UND Database*

Yale B database and the UND database, we discuss several observations and issues in this section.

In the proposed method, we used weights for different atoms which were obtained by calculating mutual information between the query sample and each of the training samples. However, the process of calculating the mutual information is time consuming. In order to reduce the time required, images were reduced in size which allowed the mutual information to be calculated faster. As previously mentioned, the l^0 solver, is much faster than l^1 solver in finding the sparsest solution. Without considering the calculation time for mutual information, our method is the fastest method for recognition among these methods as shown in Table 5.3 . However, the SL0 and SRC methods do have an advantage since there is no need to calculate mutual information. Based on these facts, We believe that by using mutual information as weights in the optimization process, the algorithm converged faster to find the global extermum, as some of the non-relevant gallery samples were discarded because of the small value of mutual information.

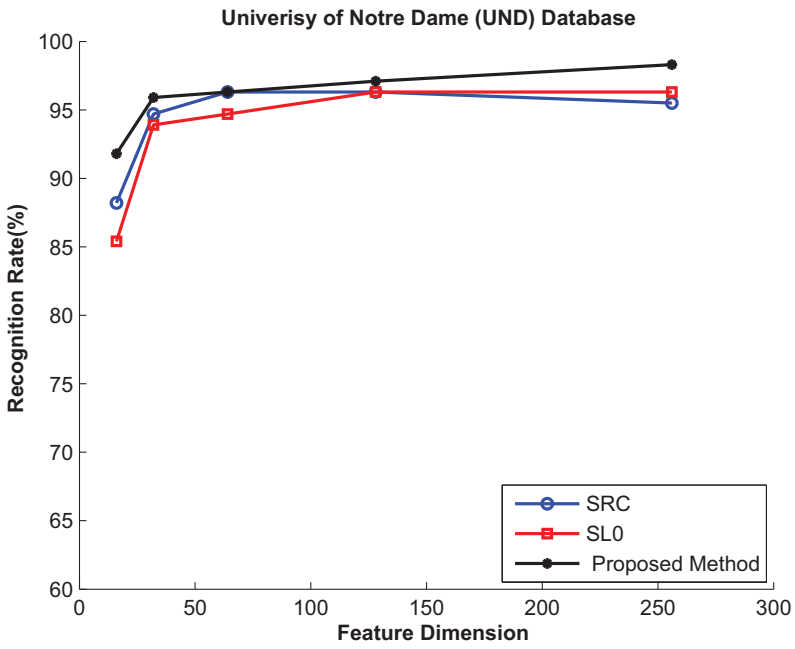


Figure 5.6: *Ear Recognition Rates on UND database Using Gabor Wavelets for Feature Extraction (10 images per subject are used for training)*

Observing the results obtained from the Extended Yale B and UND databases, our algorithm had robust results when using fewer training samples. The first experiment was conducted on the Extended Yale B database for face recognition. Our proposed method outperformed the-state-of-the-art methods for different numbers of features. In addition, to show the robustness of the proposed method with respect to the number of the training images, we varied the number of the training images from 50% to 40% and to 30% of the total number of images per subject. The results obtained by using the proposed method were far more robust than the results obtained from the other classifiers. In addition, for lower dimensional features, the recognition rate obtained by the proposed method was significantly more accurate than the other methods.

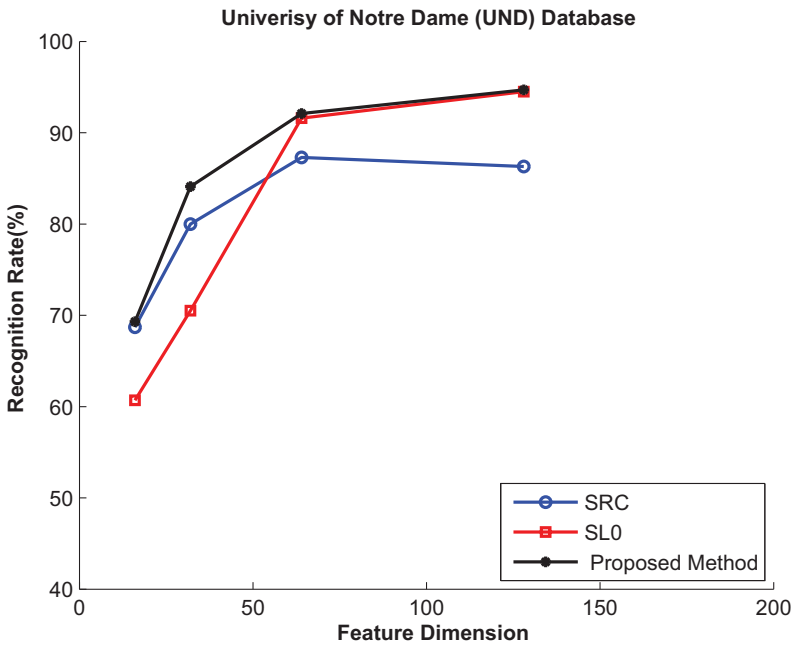


Figure 5.7: *Ear Recognition Rates on UND database Using Gabor Wavelets for Feature Extraction (5 images per subject are used for training)*

The second experiment was performed on the UND collection J2 ear database. Features were extracted using Gabor wavelets and a dictionary was constructed using the extracted features from training subjects. The subjects with 16 or more images were selected for the experiments since we needed to build a dictionary for sparse representation. Ten images per selected objects were used for training and the rest of the images were used for testing. A classification rate of 98.3% was obtained using the proposed algorithm for ear recognition, which was more precise than other classifiers. Furthermore, HOG features were used for ear recognition with different numbers of histogram bins. Experimental results showed that our proposed method, when compared to SRC, not only had a lower computation load, but also resulted in a better recognition rate.

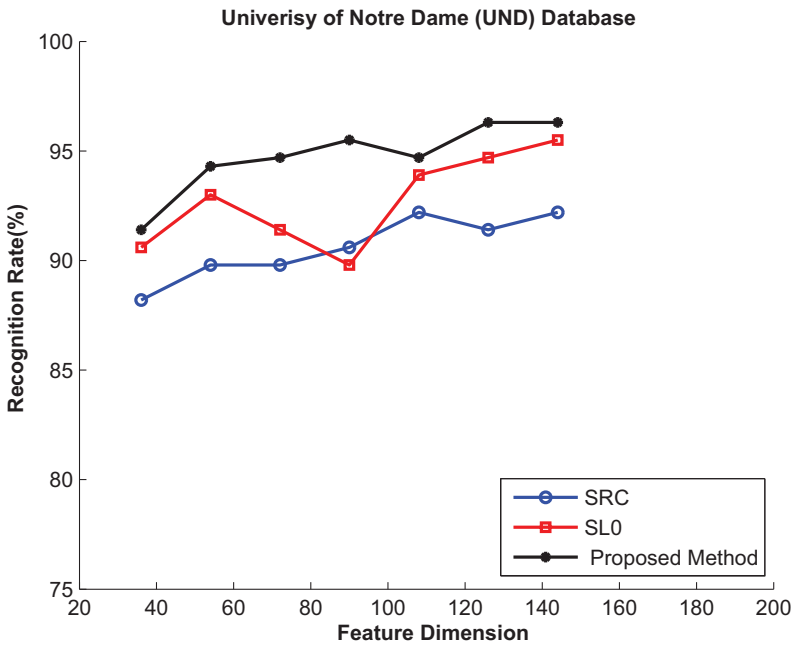


Figure 5.8: *Ear Recognition Rates on UND database Using HOG for Feature Extraction (10 images per subject are used for training)*

Finally, there is an important observation, i.e., if the values of $\|\mathbf{x}\|_0$ and $\|\mathbf{W}\mathbf{x}\|_0$ are the same, how can the weights be useful? The answer is that the SL0 algorithm uses a Gaussian function to approximate the l^0 norm, which is not the same as the exact l^0 norm. Using this approximation, the weights become significant and useful in guiding the algorithm quickly towards the solution.

Table 5.3: Time (in Seconds) for recognition of one query sample (The number of features is 56)

The Number of Training Images Per Subject	30	20	10
SRC	0.245	0.224	0.181
SL0	0.117	0.108	0.095
The Proposed Method	0.105	0.094	0.089

5.6 Conclusion

In this chapter, we presented a method for classification based on weighted sparse representation using Smoothed l^0 Norm with Non-negative Coefficients. In sparse representation classification, there is no constraint on the coefficients and all the atoms in the dictionary are treated uniformly. This means that the optimization algorithm has to search all possible solutions to converge and find the global extremum. While the SRC algorithm does not differentiate between the atoms in the dictionary, our algorithm uses mutual information to find the similarity between the query sample and samples in the gallery, which provides useful weights for more accurate classification. In fact, the obtained weights affect the significance of the atoms in the gallery during the search process. As a result, the weights narrow the search space for the algorithm and help the algorithm to converge faster without being trapped in local extremums. Furthermore, for a robust and discriminative representation, a non-negative constraint is applied on the sparse coefficient vector. By adding the non-negative constraint, components of this coefficient vector indicate the contributions

of the gallery samples towards the representation of a given query sample. We evaluated the proposed method on two publicly available databases, the Extended Yale B database for face recognition and UND database for ear recognition. Experimental results obtained for our proposed method are not only faster, but also result in a better recognition rate even with a smaller number of training samples.

CHAPTER 6

Robust Biometrics Recognition using Joint Weighted Dictionary Learning and Smoothed L0 Norm

In this chapter, we present an automated system for robust biometric recognition based upon sparse representation and dictionary learning. In sparse representation, extracted features from the training data are used to develop a dictionary. Classification is achieved by representing the extracted features of the test data as a linear combination of entries in the dictionary. Dictionary learning for sparse representation has shown to improve the results in classification and recognition tasks since class labels can be used in obtaining the atoms of learnt dictionary. We propose a joint weighted dictionary learning which simultaneously learns from a set of training samples an over complete dictionary along with weight vectors that correspond to the atoms in the learnt dictionary. The components of the weight vector associated with an atom represent the relationship between the atom and each of the classes. The weight vectors and atoms are jointly obtained during the dictionary learning. In the proposed method, a constraint is imposed on the correlation between the obtained atoms that represent different classes to decrease the similarity between these atoms. In addition, we use smoothed L0 norm which is a fast algorithm to find the sparsest

solution. Experiments conducted on the West Virginia University (WVU) and the University of Notre Dame (UND) datasets for ear recognition show that the proposed method outperforms other state-of-the-art classifiers.

6.1 Overview

Over the past few years, the theory of sparse representation has been used in various practical applications in signal processing and pattern recognition [4]. A sparse signal can be represented as a linear combination of a relatively few base elements in an over complete dictionary [80]. Sparse representation has been used for compression [12], denoising [13], and audio and image analysis [14]. Wright et. al. [20] proposed a classification algorithm for face recognition based on a sparse representation (SRC). The reported results for face recognition are encouraging enough to extend this concept to other areas such as biometrics [23]. Naseem et al. [3] addressed the problem of human identification using ear biometrics in the context of sparse representation. They used l^1 norm minimization to find the sparsest solution for ear representation. They cropped the ear portion from each image and normalized the ear region. They conducted several experiments using the University of Notre Dame (UND) database [58]. In this chapter, we use SL0 algorithm to find the sparsest solution for classification. SL0 is a fast algorithm for over complete sparse decomposition. In fact, this method finds sparse solutions for under determined systems of linear equations. Previous methods usually solve sparse problems by minimizing l^1 norm using linear programming (LP) algorithms. However, SL0 algorithm directly

minimizes the l^0 norm and it is about two to three orders of magnitude faster than state-of-the-art LP algorithms [22].

The obtained dictionary using training samples should be able to span the subspace of all samples from one subject to give a discriminative reconstruction. The basic approach to build a dictionary is to extract a feature vector for each training sample and use it as a column or atom in the dictionary. This approach is straight forward and have several disadvantage. The main disadvantage is the huge size of the dictionary. For a small database with a few number of training samples per subject, it is not a big problem. However, for a large database with thousands of subjects and many training samples per subject, the size of the dictionary becomes a serious problem. Not only there is a need for large memory to save the dictionary, but also the recognition process will be slow and not appropriate for practical applications. In addition, all the training samples may not be useful for spanning the subspace, for example if some of the training samples for a subject are similar to each other, there is no need to use all of them in the dictionary. Manual selection of a subset of training samples to construct the dictionary can not provide an optimal solution. Recently, discriminative dictionary learning has been studied in various pattern recognition and classification problems and algorithms for learning a dictionary and using less number of atoms have been developed [81], [82], [83]. One of the main methods for dictionary learning is the K-SVD method [12] which learns an over-complete dictionary and decreases the number of atoms in the dictionary. Inspired by K-SVD, many unsupervised dictionary learning algorithms have been developed and well adapted for reconstruction tasks such as restoring a noisy signal. Recent works have shown that good performance can be achieved when the dictionary is tuned to the specific task it

is intended for. Duarte et al. [84] proposed a dictionary learning method for compressive sensing, and in [85], dictionaries are developed for signal and image classification. This type of approach for dictionary learning are called task-driven algorithms [86].

In this chapter, we propose a robust recognition algorithm using sparse representation and dictionary learning which is fast and practical for real world applications and because of that we use a few atoms in the dictionary. We use a dictionary learning method to find a few representative atoms from many training samples. In fact, we try to reduce the number of atoms in the dictionary in order to decrease the processing time. WVU is used to show the effectiveness of our proposed method since it has different viewing angles for the ear and we could extract separate training and testing sets based on the viewing angles. 35 frames per subject, which approximately cover the range of the camera positions from 0 to 34 degrees, are extracted. A dictionary is obtained using Joint Weighted dictionary Learning (HWDL) which is developed with a few atoms (5, 7 or 9) for each subject to build a fast and accurate system for ear recognition. The proposed method is developed to be practical in real world applications.

This chapter is organized as follows: In Section II, we provide a brief mathematical explanation of the sparse representation concept, the proposed dictionary learning algorithm, and Smoothed L0 algorithm. In Section III, we present the experimental results to demonstrate the performance of the proposed method. Conclusions and future research directions are discussed in Section IV.

6.2 Classification based on Sparse Representation

Under determined systems of equations are important in variety of application such as signal processing, statistics, pattern recognition and image processing. Sparse representation is a relatively new approach to solve these systems. In this section, we explain the concept of sparse representation and introduce the approach for building and learning the dictionary. Finally, a brief explanation of smoothed l^0 norm (SL0) algorithm is provided.

6.2.1 Building the Dictionary

In the proposed method, a dictionary is built using the training data. The dictionary is a matrix in which each column is the feature vector of one of the training samples. Assume that there are n_i training data samples for the i^{th} class, where each data sample is represented by a vector of m elements. The matrix (dictionary) $\mathbf{A}$ is built of all the training samples from all classes as:

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_k] \in \mathbf{R}^{m \times n} \tag{6.1}$$

where k is the number of classes, $\mathbf{A}_i$ is the dictionary for class i and $n = \sum_{i=1}^{k} n_i$ and matrix $\mathbf{A}$ contains dictionaries for all the classes. A linear representation for the feature vector of the test data, $\mathbf{y}$, can then be given as:

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 \in \mathbf{R}^m \tag{6.2}$$

where $\mathbf{x}_0$ is the sparse coefficient vector. For a test data, $\mathbf{y}$, belonging to the i^{th} class, it is assumed that the non-zero elements of $\mathbf{x_0}$ will correspond to the training data

samples from the i^{th} class. However, due to noise and representation errors, there will be extraneous non-zero elements corresponding to other classes. To obtain $\mathbf{x}_0$, the equation $\mathbf{y} = \mathbf{A}\mathbf{x_0}$ should be solved such that $\mathbf{x_0}$ is sparse.

The sparsest solution of $\mathbf{y} = \mathbf{A}\mathbf{x_0}$ can be obtained by minimizing l^0 norm.

$$\widehat{\mathbf{x}}_0 = argmin\|\mathbf{x_0}\|_0 \quad Subject\ to \quad \mathbf{y} = \mathbf{A}\mathbf{x_0} \tag{6.3}$$

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

6.2.2 Sparse Solution Based on Smoothed l^0 norm Minimization

The l^0 norm of a vector is a discontinuous function and therefore it is highly sensitive to noise. In addition, combinatorial search is needed for minimizing l^0. The idea of SL0 is based on the approximation of the discontinuous function by a continuous one. This approximation is performed using a parameter (σ) which determines the quality of the approximation. Once we obtain a continuous function, we can use an optimization method, such as LevenbergMarquardt, GaussNewton or gradient descent for minimization [45].

One example for such approximations is as follows [22]:

$$f_\sigma(x) \triangleq exp(\frac{-x^2}{2\sigma^2}) \tag{6.4}$$

And approximately:

$$f_\sigma(x) \approx \begin{cases} 1, & if \quad |x| \ll \sigma \\ 0, & if \quad |x| \gg \sigma \end{cases}$$

Then, the idea is to minimize l^0 norm, $\|\mathbf{x}\|_0$, using the following function:

$$F_\sigma(\mathbf{x}) = \sum_{i=1}^{r} f_\sigma(x_i) \tag{6.5}$$

In recognition problems, r is the number of training data. Hence, we can conclude that for small values of σ, $\|\mathbf{x}\|_0 \approx r - F_\sigma(\mathbf{x})$ and to find the minimum l^0 norm solution, $F_\sigma(\mathbf{x})$ should be maximized.

Briefly, in SL0 algorithm, $F_\sigma(\mathbf{x}) \triangleq \sum_i \exp(-\mathbf{x}_i^2/2\sigma^2)$ should be maximized for a given value σ subject to $\mathbf{y} = \mathbf{A}\mathbf{x}$. A decreasing sequence of σ is used to decrease the chances of obtaining local extrema. For the initial value of σ, F_σ is maximized subject to $\mathbf{y} = \mathbf{A}\mathbf{x}$ using the steepest ascent approach. The $\mathbf{x}$ that maximizes F_σ will be the starting point to find $\mathbf{x}$ that maximizes F_σ for the next (smaller) σ.

In steepest ascent approach, each iteration moves in the desired direction ($\mathbf{x}' \leftarrow \mathbf{x} + \eta \nabla F_\sigma$), followed by projection to the feasible set $\mathcal{S} = \{\mathbf{x}|\mathbf{y} = \mathbf{A}\mathbf{x}\}$ [46]:

$$\widehat{\mathbf{x}}_0 = argmin_{\mathbf{x}} \parallel \mathbf{x} - \mathbf{x}' \parallel \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x} \tag{6.6}$$

$$= \quad \mathbf{x}' - \mathbf{A}^\dagger(\mathbf{A}\mathbf{x}' - \mathbf{y})$$

where $\mathbf{A}^\dagger = \mathbf{A}^\mathbf{T}(\mathbf{A}\mathbf{A}^\mathbf{T})^{-\mathbf{1}}$ is the pseudo-inverse of $\mathbf{A}$. Moreover, the initial value for $\mathbf{x}$ is provided by the minimum l^2 norm solution of $\mathbf{y} = \mathbf{A}\mathbf{x}$, that is, $\mathbf{A}^\dagger\mathbf{y}$.

6.2.3 Classification

In SRC, classification is based upon the obtained $\mathbf{x}$ by computing the error between $\mathbf{y}$, the original test data, and $\widehat{\mathbf{y}}_\mathbf{i}$, the approximation obtained through the sparse representation. For each class i and $\mathbf{x}_i \in R^n$, the vector $\delta_i(\mathbf{x}_i) \in R^n$ contains only the coefficients that are associated with class i and zeros for the coefficients associated

with the other classes. Using this definition, approximated test data $\widehat{\mathbf{y}}_\mathbf{i}$ is computed as follows:

$$\widehat{\mathbf{y}}_\mathbf{i} = \mathbf{A}\delta_i(\mathbf{x_i}) \tag{6.7}$$

classification can subsequently be performed by assigning the test data to the class that minimizes the residual between $\mathbf{y}$ and $\widehat{\mathbf{y}}_\mathbf{i}$ as follows:

$$\underbrace{min}_{i} \quad r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\delta_\mathbf{i}(\mathbf{x_i})\|_\mathbf{2} \tag{6.8}$$

where $r_i(\mathbf{y})$ is the residual distance for class i.

6.2.4 Joint Weighted Dictionary Learning

As previously mentioned, the matrix $\mathbf{A}$ may have a huge size which makes the recognition process time consuming. K-SVD method [12] has been used for dictionary learning, which is an iterative approach that alternates between sparse coding of the atoms in the current dictionary and updating the dictionary for more discriminative representation. Most of dictionary learning techniques and data driven methods such as the K-SVD consider a finite training set of samples and minimize the empirical cost function. The K-SVD algorithm finds the solution for the following problem:

$$\underbrace{argmin}_{\mathbf{D},\mathbf{Y}} \quad \|\mathbf{A} - \mathbf{DY}\|_F^2 \quad subject\ to \quad \forall_i, \|\mathbf{y_i}\|_0 \leq P \tag{6.9}$$

where P is a parameter that defines the required sparsity, and $\mathbf{D}$ is the learned dictionary with a smaller number of atoms than $\mathbf{A}$. The non-convex optimization problem of Eq.6.9 can be iteratively solved by fixing one parameter such as $\mathbf{D}$ that makes it a convex optimization problem with the other parameter as $\mathbf{Y}$. After finding the $\mathbf{Y}$, it will be the fixed parameter and we solve the problem to obtain the $\mathbf{D}$.

The optimization of Eq.6.9 is unsupervised in the sense that it does not require the use of the labels for the atoms. However, in this paper, we introduce a dictionary learning method which is developed for specific supervised tasks, e.g., classification or recognition, as opposed to the unsupervised formulation of the data driven methods. In classification applications, a good data representation can lead to an accurate performance and our method improves the representation by learning more efficient and discriminative atoms for sparse representation. In the proposed method, instead of learning the dictionary without considering the labels, we use the training samples of each class for learning the atoms of the dictionary and finding the related weights. In fact, this method helps to learn a discriminative dictionary $\mathbf{D} = [\mathbf{d_1}, \mathbf{d_2}, ..., \mathbf{d_N}]$, where N is the number of atoms in the learned dictionary, and obtains atom weights that can be included in a weight matrix $\mathbf{W} = [\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_k}]$, which indicate the relationship between atoms and the classes. The $\mathbf{w}_i = [w_{i,1}, w_{i,2}, ..., w_{i,N}]^T$ indicates the weight vector between each atom of learned dictionary $\mathbf{D}$ and the i^{th} class as shown if Fig 6.1. The goal is to find $\mathbf{D}, \mathbf{w}_i$ and $\mathbf{Y}_i$ for $\mathbf{A}_i \approx \mathbf{D}diag(\mathbf{w}_i)\mathbf{Y}_i$, that can well represent the class i dictionary, $\mathbf{A}_i$. It is worth mentioning that the weight vector helps the learnt dictionary to well represent all training samples in i^{th} class. Non-negativity constraint is imposed on the weights elements, $w_{i,m} \geqslant 0, \forall m$ as there is no negative relation between an atom and class. If one atom can not represent a class or there is no relation between that atom and the class, the associated weight to that class will be zero. The sum of all weight elements for one atom is normalized to one as $\sum_m w_{i,m} = 1, \forall m$. Thus, we arrive at the following joint weighted dictionary

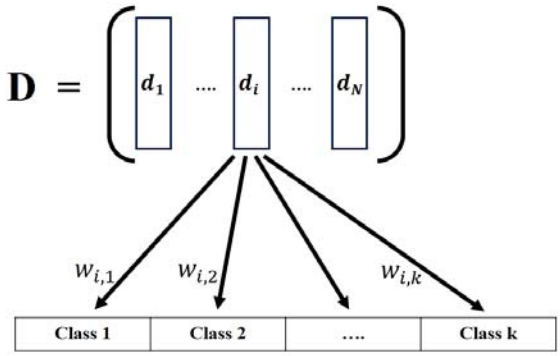


Figure 6.1: *The components of the weight vector associated with an atom represent the relationship between the atom and each of the classes*

leaning model:

$$\underbrace{argmin}_{\mathbf{D},\mathbf{W},\mathbf{Y}} \quad \sum_{i=1}^{k} \parallel \mathbf{A}_i - \mathbf{D}diag(\mathbf{w}_i)\mathbf{Y}_i \parallel_F^2 + \lambda_1 \parallel Y_i \parallel_1$$

$$+\lambda_2 \sum_{i=1}^{k} \sum_{l \neq i} \sum_{n=1}^{N} \sum_{m \neq n} w_{i,m}(\mathbf{d}_m^T \mathbf{d}_n)^2 w_{l,n} \tag{6.10}$$

$$s.t. \quad w_{i,m} \geqslant 0 \quad and \quad \sum_{m} w_{i,m} = 1, \forall m$$

where k is the number of classes, $\mathbf{Y}_i$ is the sub-matrix which has the sparse coefficients of $\mathbf{A}_i$ over $\mathbf{D}$. In this equation, the discrimination is exploited using the dictionary itself and the sparse coefficients associated to $\mathbf{D}$. The term $(\mathbf{d}_m^T \mathbf{d}_n)$ in

$$\sum_{i=1}^{k} \sum_{l \neq i} \sum_{n=1}^{N} \sum_{m \neq n} w_{i,m}(\mathbf{d}_m^T \mathbf{d}_n)^2 w_{l,n}$$

in Eq.6.10 is the correlation between two atoms. In fact, this term is added so that if two atoms (d_m and d_n) in the dictionary are very similar to each other, the weights ($w_{i,m}$ and $w_{l,n}$) will become smaller. It is obvious that we need more discriminative atoms instead of similar atoms in the learned dictionary in order to represent a query sample. The obtained dictionary $\mathbf{D}$ and the weight matrix $\mathbf{W}$ can more accurately

represent a query sample as the atoms are learned optimally to well represent each class individually.

6.2.5 Feature Extraction

We use Histogram of Oriented Gradients (HOG) descriptor for ear recognition, which was first presented and efficiently used for object detection and image retrieval [76], especially when illumination variations are present. Actually, it is considered as one of the best features for the dense encoding of 2D image regions, and has been successfully used in pedestrian detection and object classification tasks [77]. HOG feature extraction is used for ear recognition since the research has shown that the HOG is one of the best features to capture the local shape information. It was also demonstrated that it achieves excellent performance in image retrieval [76] and 2D object detection tasks [77]. The HOG feature descriptor is in fact a dense version of the SIFT feature descriptor, i.e., SIFT descriptor computed on a dense grid. For building a feature vector using HOG, the cell histogram of each pixel within the cell casts a weighted vote based on the gradient l^2 norm, for an orientation-based histogram channel. The gradient strength was locally normalized in order to account for changes in illumination and contrast [78].

6.3 Experiments

In this section, several experiments are performed to demonstrate the effectiveness of the proposed method in terms of recognition accuracy. We present our experimental results on two publicly available databases for ear recognition. We evaluate our

method using HOG features and compare it with other classification algorithms such as Nearest Neighbour (NN) and Nearest Subspace (NS), as well as the original SRC and SL0. In addition, two other classification algorithms based on sparse representation, IRL1 [87] and SWSR-COS [88], are used for comparing results.

The University of Notre Dame (UND) dataset collection J2 is used to test the proposed method for 2D ear recognition. The Histograms of Indexed Shapes (HIS), a shape-based feature set, is used to localize a rectangular region that contains the ear [65], [54].

We used an equal number of images (10 images per subject) from each subject for training and the remaining images were used for testing. HOG was used for feature extraction and the number of features were reduced using PCA. As mentioned previously, the equation $\mathbf{y} = \mathbf{Ax}$ should be under determined and the number of columns in the dictionary should be more than the number of rows.

The number of features are reduced to 16, 32, 64, 128 and 256 using PCA. Fig. 6.2 shows the comparison of the recognition results of our algorithm with SRC, SL0 and SL0+KSVD when using different number of features. The obtained results using our proposed method show significant improvement in the recognition accuracy.

Here, we describe the second experiment that we performed in order to evaluate the proposed approach and present the results. We present experiments for ear recognition using the WVU data set, which consists of video sequences captured by a rotating camera around the head of different subjects. The ear region is extracted in each image automatically using proposed algorithm in [57] which uses a shape based feature set, termed the Histogram of Indexed Shapes (HIS), to localize a rectangular region that contains the ear region. The video sequences start from the left profile

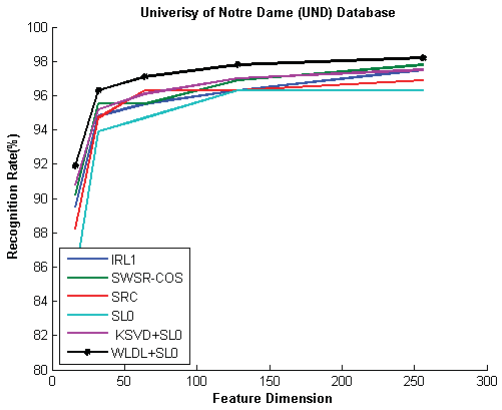


Figure 6.2: *Ear Recognition Rates on UND database (10 images per subject are used for training)*

Figure 6.3: *A few samples of extracted frames for one subject for different viewing angles*

of each subject (0 degrees) and terminate at the right profile (180 degrees) [68]. The length of each video sequence is about two minutes. A few subjects in the data set have eyeglass, earrings or part of the ear is occluded by hair. There are three subjects that have their ear fully occluded and these subjects were not used in the experiment.

In this experiment, 35 frames, which approximately cover the range of the camera positions from 0 to 34 degrees (*i.e.*, one frame for each degree) from 100 subjects, are extracted. Fig. 6.3 shows a few samples from a video sequence for one of the subjects. After extracting the frames, the ear region is detected and a bounding box

around the ear is extracted. The ear detection is performed automatically based on the algorithm in [57]. Since the sizes of the extracted bounding boxes vary, we normalized the size to 120x80. We use 20 frames for training and 15 frames for testing from each subject. The joint weights dictionary learning is used to learn a dictionary with 5, 7 and 9 atoms for each subject and for SRC and NN, 5, 7 and 9 images of each subject are selected randomly. The length of the feature vector is 32. The experiment is performed 10 times and average results are shown in Table7.1. It is obvious from obtained results that the proposed method outperforms other approaches. For example, for 9 atoms in the dictionary, NN performs good by 79% better than SVM and Adaboost. However, dictionary learning methods, K-SVD and JWDL, outperform other classifiers by 82% and 85% accuracy. It shows that our proposed method is far better than other approaches. Furthermore, we decreased the number of features from 32 to 16, and and repeated the experiment. Similar to the previous experiment, the dictionary is learned for 5, 7 and 9 atoms and for SRC and NN the same number of images were selected randomly. The obtained results are shown in Table7.2 , and again are consistent with the results shown in Table 1.

6.4 Conclusion

In this paper, we presented a new method for biometrics recognition based on sparse representation using Smoothed l^0 Norm and joint weighted dictionary learning. Classification is achieved by representing the query sample as linear combination of the training samples. Joint weighted dictionary learning simultaneously learns from a set of training samples of overcomplete dictionary along with weight vectors that

Table 6.1: Ear Recognition Rates on WVU database (feature vector size is 32)

Number of Atoms in Dictionary	5	7	9
NN	74.1%	77.3%	79.8%
SVM	72.8%	74.1%	76.5%
Adaboost	68.9%	72.3%	75.5%
SRC	65.1%	66.0%	68.2%
K-SVD + SL0	77.5%	78.1%	82.1%
JWDL + SL0	78.8%	81.1%	85.5%

associated to the atoms in the learnt dictionary. In fact, we define the relationship between each atom and all the classes by weight vector. The weight vectors and atoms are jointly obtained during the dictionary learning. Smoothed L0 norm is used to obtain the sparsest solution. We evaluated the proposed method on two databases, UND and WVU databases. Usually in practical problems, the training images of the subject are captured at certain angles, which are different from the angles for the test images. In this paper, we trained and tested the system with images captured at different viewing angles to mimic what happens in practical applications. Experimental results show that the proposed method is not only faster than previous methods, but also has a better recognition rate even with a smaller number of training samples.

Table 6.2: Ear Recognition Rates on WVU database (feature vector size is 16)

Number of Atoms in Dictionary	5	7	9
NN	74.5%	76.6%	78.1%
SVM	71.5%	73.8%	74.3%
Adaboost	68.9%	70.3%	71.5%
SRC	66.2%	66.6%	68.8%
K-SVD + SL0	76.8%	76.6%	80.5%
JWDL + SL0	77.5%	78.9%	83.1%

CHAPTER 7

Robust Object Tracking via Adaptive Sparse Representation and Feedback

We are going to present a tracking system based upon adaptive sparse representation and dictionary learning. Developing an effective and complete tracking algorithm is a challenging task because of factors such as illumination, occlusion and pose variations. Most of the tracking algorithms do not consider the situation when the tracked object or disappears temporarily from the video sequence or becomes temporarily fully occluded. In this chapter, our goal is to develop an automatic object tracking system that can handle pose variations, scale variations and temporary disappearance of the object from the scene. We present a robust tracking system based on adaptive sparse representation and feedback. We focus on automatic tracking with no prior knowledge other than the location of the region to be tracked in the first frame, which can either be located manually or using a detector that finds the region of interest (ROI). The visual tracking is a binary classification problem. The positive samples are bounding boxes that have high overlap with current position of the target while negative samples are drawn from regions outside the ROI to model background close to the target. The tracking algorithm uses the dictionary to locate

the ROI in the following frames via adaptive sparse representation. One of the main issues in tracking systems is false tracking when the object disappears from the scene. Motivated by the concept of feedback in control systems, we overcome the problem of false tracking when the object disappears by comparing the newly tracked region with previous regions to confirm that the object is still in the frame. A structural similarity measure is used to measure similarity between a newly tracked ROI and the previously tracked ROIs and if the similarity is below a certain threshold, the object is assumed to be out of the scene. In fact, this similarity evaluation is like a feedback loop in our tracking algorithm which makes our method robust, reliable and accurate when compared to the state-of-the-art methods on challenging sequences. If the object is not located in the current frame, the algorithm stops tracking and starts searching for the object in the following frames. The searching is achieved by using a detector based on sparse representation and an adaptive dictionary to efficiently locate the object when it reappears in the scene. Experiments on video sequences, for both quantitative and qualitative evaluations, demonstrate the effectiveness and robustness of the proposed tracking system.

7.1 Overview

Object tracking is a well studied problem in computer vision and it has many practical applications [89], [90], [91]. Tracking is a very challenging task since geometric and photometric factors, *e.g.*, occlusion, pose and illumination vary that can lead to change the appearance of the object [92], [93], [94] and therefore could cause tracking errors. Furthermore, severe motion blur usually occur when a video has a

low frame rate or when an object moves abruptly. Although there has been some success with tracking methods for specific object classes, *e.g.*, faces [95], humans [96], rigid objects [97], tracking generic objects is still a challenging and interesting problem for researchers. A typical tracking system has three main components [98]: 1) an appearance model or object representation which measures the likelihood of the object to be at a certain location in the frame, 2) a motion model, which relates the location of the target through the video sequence, and 3) a search method to find the most likely location of the ROI in the current frame. During the last decade, supervised methods have been proposed to solve these problems [99], [100]. Most of these methods either use a pre trainned classifier for a specific object or train a classifier using the selected bounding box in the first frame. The classifier is used and updated for object tracking in the following frames. However, when object is occluded or moves out of the field of view may drift these trackers. If the object disappears, some negative samples may falsely be classified as positive samples, which mislead the classifier and gradually undermine the model [101].

It is well established that sparse signal models are well suited for recognition and classification tasks and can be effectively learned from audio and image data. Sparse representation has received a widespread attention because of its robust performance and wide range of applications. During the last decade, the theory of sparse representation has been used in various practical applications in signal processing, pattern recognition, and video analysis [4], [8], [7], [9]. It has also been used for compression [12], denoising [13], and audio and image analysis [14]. In addition, dictionary learning and sparse representation have been used as powerful tools for classification and analysis of image and video data [102], [103]. Generally, sparse representation

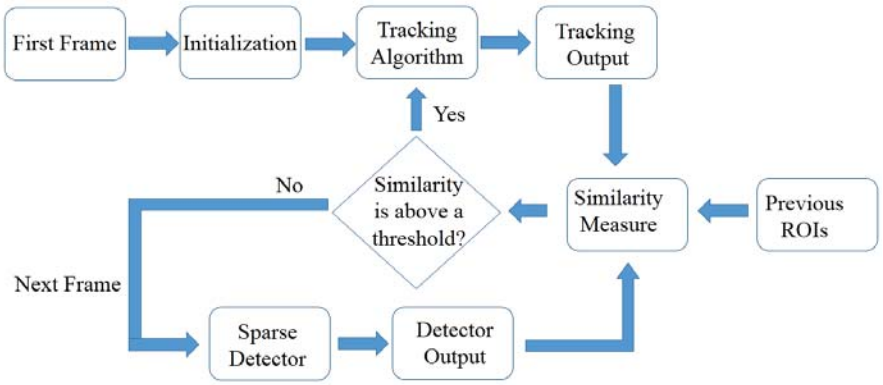


Figure 7.1: *The flowchart for the proposed tracking system.*

is a technique for reconstructing a signal or image using the fact that signals can be represented by a set of basis elements [18]. Sparse representation uses all the data samples for the decision making and represents the test data as a linear combination of the training data. Mei and Ling [104] proposed a visual tracking method by formulating the problem as a sparse approximation problem in a particle filter framework. Chen *et al.* [105] presented an appearance based tracking method using sparse representation. In their method, the appearance of an object is modelled by multiple linear subspaces.

In this paper, we propose a robust tracking system based on sparse representation by posing the tracking as a classification problem. The main idea of our method is to utilize sparse representation and similarity measure to efficiently construct a robust tracking algorithm which can detect the object after it temporarily disappears or becomes temporarily occluded. The flowchart of the proposed system is shown in Fig. 7.1. The First step is initialization, *i.e.*, choosing the tracked target or region of

interest (ROI) either manually or using an automatic detector. To build and update the dictionary for sparse representation, we need positive and negative samples in each frame. Positive and negative samples for one frame are shown in Fig. 7.2. For object representation, we need an over complete dictionary with labeled data from the first frame. Both positive and negative samples are used to build the dictionary. Most of the trackers drift from tracking the target when target is occluded for long time. We propose to solve this issue by comparing the estimated location of the target from tracker with the obtained template of the target. This way we can make sure that the target candidate has standard similarity to the target template and it is not false positive. If the object is not in the frame, the algorithm attempts to detect the object in the following frames using sparse representation and resumes tracking after detecting the object.

The remainder of this paper is organized as follows: In Section 2, we review the current state of the art in tracking systems, sparse representation and detection; In addition, the concept of feedback in control systems is described. In section 3, we detail our tracking approach; in Section 4, qualitative and quantitative results of our tracker and state of the art trackers on video clips that include publicly available video sequences are presented. Finally, we conclude the paper in Section 5.

7.2 Related Work

This section reviews the related methods for appearance model-based tracking, compressed sensing, object detection and image similarity measurements.

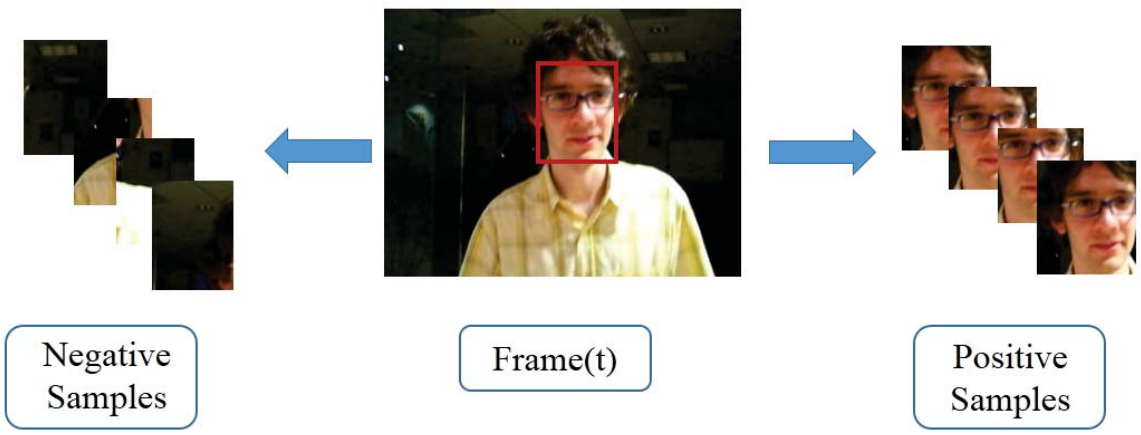


Figure 7.2: *Sample set in frame t, Left: Negative Samples, Right: Positive Samples*

There are studies that treat tracking as a classification or representation problem where learning techniques are used to locate the object in the sequence of frames [106]. A good representation should provide a strong description to distinguish the target from the background and other objects. To account for the appearance variations of the target during tracking, many sophisticated object representation methods have been developed including, generative, discriminative and sparse representation methods [107], [108], [102]. The generative model-based approaches describe the visual observations of a moving object. Hence, tracking is reduced to a search for an optimal state that yields an object appearance most similar to the appearance model. On the other hand, the discriminative methods use a binary classification, based on an optimal decision boundary, to distinguish the object from the background [109]. For generative appearance modeling methods, a tracking method was presented in [110] based on subspace constancy assumption. A subspace model was constructed offline from some collected target observations where the model was fixed during tracking. This is a big drawback as for most tracking applications, it might be difficult to ob-

tain such target observations in advance. Consequently, their method has limited application domains and may fail when the target shows at a different view from the ones used for constructing the model. For discriminative appearance modeling methods, an online feature selection method was proposed in [111] to select the most discriminative color spaces for tracking. Unlike the generative models which only model the target, the discriminative models can model both the target and background. However, they need correctly labeled samples to train and update classifiers, which may not be available in many real tracking applications [112]. On the other hand, sparse representation can recover the appearance subspace when corruption (noise and occlusion) of the observation is large but sparse. The global appearance of a target in different conditions can be well represented by a linear combination of a small number of training images in an over-complete image dictionary. This representation has been successfully used in face recognition [113] and ear recognition [23] applications. In our method, sparse representation is used for tracking an object by representing a candidate as a linear combination of the previously tracked ROIs of the object over the previous frames. Basically, the tracking task is formulated as a classification problem with online update of the dictionary.

Compressed sensing is an efficient data acquisition method whenever the data is sparse in a high dimensional space. Basically, a signal that is sparse in a known transform domain can be represented with much fewer samples than usually required by the dimensions of this domain. In fact, it is shown that if the dimension of the feature space is sufficiently high, these features can be projected to a randomly chosen low-dimensional space which contains enough information to represent the original high-dimension features [80]. From the machine learning point of view, compressed

sensing can be regarded as efficient universal sparse dimensionality reduction from the data domain to the measurement domain [114]. There are applications where compressed sensing is used for feature reduction, for instant face detection [115]and natural language processing and text classification [116]. In the proposed method, we use compressed sensing for feature reduction using a very sparse measurement matrix that asymptotically satisfies the restricted isometry property (RIP) in compressed sensing theory [117] which provides an efficient projection from the image feature space to a low-dimensional compressed subspace [118]. This dimensionality reduction speeds up our method for fast and online tracking.

Visual tracking can be combined with either detection [119], [120] or recognition [121], [122]. Object detection is the task of finding and locating objects in an input image or video frame. The definition of an "object" can be a single instance or a whole class of objects. In this paper, we define object as a single instance that may have rotation or scale changes. There are three main components for feature based approaches: 1) feature detection, 2) feature recognition, and 3) model fitting. Vacchetti *et al.* [123] presented a tracking algorithm for rigid objects in 3D using a single camera that can handle large camera displacements and partial occlusions. This tracker combines natural feature matching and the use of key frames to handle camera displacement. Taylor and Drummond [124] proposed a feature matching tracker which enables seven independent targets to be localized in a video sequence using histogrammed intensity patches (HIPs). The main draw back for these approaches is the detection of image features and requirement of knowing the geometry of the object in advance. On the other hand, the sliding window-based methods scan the input image by a window of various sizes and each scanned window is classified as either

having the object (ROI) or not. Viola and Jones [125] presented an object detection which used integral image, Adaboost and cascade classification for detection. Sliding window based detectors can be time consuming. In the cascade architecture of Viola and Johns, since the background usually has more variations than the object, the classifier is separated into number of stages to enable quick rejection of background patches and reducing the number of stages that have to be evaluated on average. For these detectors, it is necessary to have a large number of training examples and intensive computation in the training stage to accurately represent the decision boundary between the object and background. In addition, training samples should be labeled.

Image similarity indices are crucial in the development and evaluation of image processing and pattern recognition algorithms such as image coding, denoising, segmentation, registration and recognition [126]. Each image is a 2-D function, $x(i, j)$, of intensity. To calculate a similarity index for images, intensity variations and geometric distortions should be accounted for. Similarity indices algorithms can be classified according to their approach toward these two properties. Some algorithms compare images based on the assumption that image are at the same scale and are perfectly registered. Therefore, their similarity is calculated from a comparison of the corresponding pixel intensities and are called intensity-based. On the other hand, geometry-based indices are determined by establishing pixel correspondences between the images based on the intensity and then similarity is calculated by comparing the geometric transformations between corresponding pixels. In intensity-based indices, the similarity evaluation at one pixel is independent of all other pixels in the image. However, we know that the neighboing pixels can be correlated with each other. Variety of transformed-domain algorithms have been proposed to take advantage of

this correlation and also to take into account properties of the human visual system (HVS) [127]. In this paper, we use a similarity index named structural similarity index measurement (SSIM) [2] to compare the newly tracked ROI with previously tracked ROIs. In SSIM, the structural information of an image means the attributes that represent the structure of the object in the visual scene, apart from the intensity and contrast. Therefore, SSIM compares local patterns of pixel intensities which have been normalized for mean intensity and contrast. SSIM has low computational complexity and robust performance in image similarity comparisons.

Despite sharing ideas with previous work, as discussed above, to the best of our knowledge, our tracking system is the first that completely addresses the disappearance of the tracked object from the field of view. We propose the use of feedback, as in control systems to enhance the performance of our algorithm. The goal of any control system is to measure, monitor, and control a process. One way in which we can accurately control the process is by monitoring its output and feeding it back to compare the actual output with the desired output. In fact, the output of the system is brought to the original or desired response. The measure of the output is called the feedback signal and the type of control system which uses feedback signal to control itself is called a closed-loop system or feedback control system and consists of controller, system and sensor. Just to clarify the concept of feedback, a basic feedback structure is shown in Fig 7.3. Feedback loop is the most important part of any self-regulating system. In our algorithm, we use the concept of feedback. After tracking and finding the new ROI, we calculate the similarity measure between the newly tracked ROI and the previously tracked ROIs and if the similarity is above a threshold, we continue tracking. Otherwise, we stop tracking and use the proposed

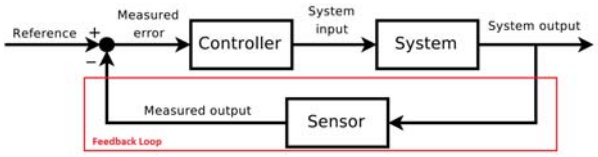


Figure 7.3: *Closed-loop feedback control system*

object detector to find the ROI in the following frames until we find the ROI and then resume tracking.

7.3 Sparse Representation based Tracking

We propose a visual tracking algorithm using classification based on adaptive sparse representation. The tracking process starts by constructing a dictionary from positive and negative samples of the object in the first frame. Then, the tracking process locates the tracked object in the following frames using sparse representation and feedback. However, the appearances of both the target and the background are likely to change because of numerous factors. To successfully track the target over time with presence of appearance changes, we need to update the observation model, *i.e.*, dictionary, when new tracking results are obtained. However, we do not update the whole dictionary since the newly tracking results may contain noise or occlusion and it can affect the representation power of our model. Instead, we only update the part of the dictionary that is most similar to the current observation.

7.3.1 Adaptive Sparse Representation

As Fig. 7.1 shows, the tracking system starts by identifying the region to be tracked either manually or using an object detector, then starts tracking the selected region in the following frames. If the region to be tracked is in frame t, the center of the bounding box, c_t, is determined and positive and negative samples are selected to build the dictionary. Positive samples are obtained in a circular area which satisfies $\| c_p - c(t) \| < \alpha$, and we draw negative samples from an annular area $\alpha < \| c_n - c(t) \| < \beta$, where c_p and c_n are the locations of positive and negative samples, respectively, and α and β are thresholds to determine the circle and annular area, respectively. The positive samples cover whole or part of the object while the negative samples should cover the background around the target location. Haar features are extracted from each sample (positive or negative). A dictionary is built by concatenation of obtained features. Each feature is a column vector of the dictionary.

The representation of an object under different illumination conditions and viewing angles is known to lie approximately in a low-dimensional subspace [128]. We use k positive samples and k negative samples, where each sample is represented by a vector of m elements. The vectors that represent the positive samples are then used to construct the columns of matrix $\mathbf{A}_p$:

$$\mathbf{A}_p(t) = [\mathbf{v}_{p,1}(t), \mathbf{v}_{p,2}(t), ..., \mathbf{v}_{p,k}(t)] \in \mathbf{R}^{m \times k} \tag{7.1}$$

where $\mathbf{v}_{p,j}(t)$, $j = 1, ..., k$, is a column vector that represents the features extracted from training sample j in frame t. Similar to $\mathbf{A}_p(t)$, matrix $\mathbf{A}_n(t)$ is built from the negative samples. The two matrices, $\mathbf{A}_p(t)$ and $\mathbf{A}_n(t)$, are concatenated to form matrix $\mathbf{A}(t)$.

In frame t, the center point of the ROI bounding box is c_t. In frame $t+1$, candidate windows are selected around center point,c_t, as candidate matches for the ROI and feature vectors are extracted from each of these samples. The goal is to determine which of these samples is the best match for the tracked ROI.

The feature vector $\mathbf{y}(t+1)$ of a sample at frame $t+1$ can be expressed as a linear combination of the entries in the dictionary:

$$\mathbf{y}(t+1) \approx \mathbf{A}(t)\mathbf{z}(t) \tag{7.2}$$

where $\mathbf{z}$ is a coefficient vector and $\mathbf{A}(t)$ is the dictionary. In visual tracking, objects are often corrupted by noise or occluded or both which create unpredicted errors. This may affect any part of the object and the size may vary. The occlusion can be either a connected region of pixels or a number of pixels that are scattered on the object. In our algorithm, all the pixels within the bounding box are treated equally for simplicity. Therefore, to handle the effect of noise and occlusion, equation (7.2) is rewritten as:

$$\mathbf{y}(t+1) = \mathbf{A}(t)\mathbf{z}(t) + \mathbf{e} \tag{7.3}$$

where $\mathbf{e}$ is the error vector, $\mathbf{e} = (e_1, e_2, ..., e_{2m})^T \in R^{2m}$ and the non-zero entries of e indicate the pixels in y which are corrupted or occluded. As the occurrence of error is random and unknown, trivial templates $\mathbf{I} = [\mathbf{i}_1, \mathbf{i}_2, ..., \mathbf{i}_d] \in R^{2m \times 2m}$ [18], where I is an identity matrix, are used to capture the occlusion and noise as:

$$\mathbf{y}(t+1) = [\mathbf{A}(t), \mathbf{I}] \begin{bmatrix} \mathbf{z}(t) \\ \\ \mathbf{e} \end{bmatrix} \tag{7.4}$$

where a trivial templae $\mathbf{i}_i$ is a vector with only one nonzero element.

7.3.2 Non-negative Coefficients

The coefficients in z can be any real numbers. Without constraint on the coefficients, the sample at frame $t+1$ can be reconstructed by adding and subtracting contributions from the training samples, as the coefficients can be positive, negative or zero. This is contradictory to the intuitive notion of combining parts to form a whole [74], [19]. In fact, using non-negative coefficients can provide insight into the similarity between the candidate sample and each of the training samples. With the constraint that the coefficients, *i.e.*, components of $\mathbf{z}$, are non negative, the coefficients represent the contributions of the individual samples in constructing the candidate sample. While we enforce the non-negativity constraint on the elements of $\mathbf{z}$, it is unreasonable to enforce that constraint on the elements of the error vector. Equation (7.4) is rewritten as:

$$\mathbf{y}(t+1) = [\mathbf{A}(t), \mathbf{I}, -\mathbf{I}] \begin{bmatrix} \mathbf{z}(t) \\ \\ \mathbf{e}_P \\ \\ \mathbf{e}_N \end{bmatrix} \doteq \mathbf{B}(t)\mathbf{x}(t), \quad s.t. \quad \mathbf{x} \succeq 0 \tag{7.5}$$

where $\mathbf{e}_P$ and $\mathbf{e}_N$ are positive and negative trivial vectors, respectively, $\mathbf{B}(t) = [\mathbf{A}(t), \mathbf{I}, -\mathbf{I}]$; and $\mathbf{x}(t) = [\mathbf{z}(t), \mathbf{e}_P, \mathbf{e}_N]^T$ is a non-negative coefficient vector with positive or zero elements. In Fig. 7.4, example samples and trivial samples are shown.

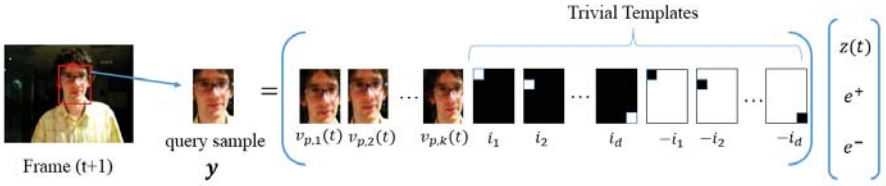


Figure 7.4: *Representation of a query sample in frame (t+1) using example samples and trivial samples*

7.3.3 Finding sparse coefficients using SL0 algorithm

Equation (7.5) is underdetermind and does not have a unique solution for $\mathbf{z(t)}$. Based on sparse representation theory, $\mathbf{z}(t)$ should be sparse since the object can be represented by a linear combination of a few atoms in the dictionary. In addition, for a good matching object candidate, there are only few non zero coefficients in $\mathbf{e}_P$ and $\mathbf{e}_N$ that account for the noise and occlusion. To obtain $\mathbf{z}(t)$, the above equation should be solved such that $\mathbf{z}(t)$ is sparse, where the sparsest solution can be obtained by minimizing the l^0 norm as follows:

$$\widehat{\mathbf{x}}_0(t) = argmin\|\mathbf{x}(t)\|_0$$

$$s.t. \quad \mathbf{y}(t+1) = \mathbf{B}(t)\mathbf{x}(t), \quad \mathbf{x}(t) \geq 0 \tag{7.6}$$

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

The l^0 norm of a vector is the number of non zero elements in the vector, which is a discontinuous function and therefore is highly sensitive to noise. In addition, combinatorial search is needed for minimizing l^0. Fortunately, the Smoothed l^0 (SL0) algorithm can be used to solve equation 7.6. The SL0 method is more efficient than the l^0 and l^1-norm minimization in term of computational complexity [22]. The idea

of SL0 is based on the approximation of the discontinuous l^0 norm function with a continuous function. This approximation is performed using a parameter (σ) which determines the quality of the approximation. Once we obtain a continuous function, it is possible to use convex optimization methods, such as Levenberg-Marquardt, GaussNewton or gradient descent for minimization [45].

One example of such approximations uses the following function:

$$f_\sigma(x) \triangleq exp(\frac{-x^2}{2\sigma^2}) \tag{7.7}$$

where approximately:

$$f_\sigma(x) \approx \{ \begin{array}{lll} 1, & if & |x| \ll \sigma \\ 0, & if & |x| \gg \sigma \end{array}$$

Then, the idea is to minimize l^0 norm, $\|\mathbf{x}\|_0$, using the following approximate function:

$$F_\sigma(\mathbf{x}) = \sum_{i=1}^{2m} f_\sigma(x_i) \tag{7.8}$$

In recognition problems, $2m$ is the total number of training samples. Hence, it is obvious that for small values of σ, $\|\mathbf{x}\|_0 \approx 2m - F_\sigma(\mathbf{x})$ and to find the solution that minimizes l^0 norm, $F_\sigma(\mathbf{x})$ should be maximized. Since $F_\sigma(\mathbf{x})$ is a continuous function, we use steepest ascent method. Each candidate sample in frame $t+1$ is classified as either the tracked object or not, based on its sparse representation. If non of the samples is classified as the tracked object, we use a detector based on sparse representation to locate the object in the following frames.

7.3.4 Similarity Measure

It is possible that the tracking algorithm loses the tracked region of interest (ROI) if the object becomes fully occluded or moves out of the field of view of the camera.

Most of the tracking algorithms continue tracking in these cases which causes drifting or updating the tracker parameters with false samples. In our algorithm, a fully occluded object can lead to errors in updating the dictionary with false samples and cause the tracker to lose object. However, we use feedback to compare the newly tracked region with the previous ones and decide if the tracked region represents the object or not. This is achieved by measuring the similarity between the newly tracked region and previous regions and if it is below a threshold, the tracking stops and our algorithm uses a detector based on sparse representation to locate the ROI in the following frames. This prevents the tracking algorithm from following a false positive instead of the correct ROI.

To measure the similarity between two images, subjective and objective methods have been presented in the literature. The structural similarity index measure was developed to compare local patterns of pixel intensities after normalizing for luminance and contrast [2]. The luminance of the surface of an object is the result of the illumination and reflectance. However, the structure of the object is independent of the illumination. Hence, to find the structural information in an image, we have to separate the influence of the illumination. In fact, the structural information in an image is defined as those attributes that represent the structure of objects in the scene, independent of the average luminance and contrast. We use the SSIM described by the system diagram in Fig. 7.5. Suppose $\mathbf{X}$ and $\mathbf{Y}$ are the previously tracked ROI and the new candidate ROI, respectively. This similarity measure is defined as a function of three components that compare luminance, contrast and structure, respectively:

$$S(\mathbf{X}, \mathbf{Y}) = f(l(\mathbf{X}, \mathbf{Y}), c(\mathbf{X}, \mathbf{Y}), s(\mathbf{X}, \mathbf{Y})) \tag{7.9}$$

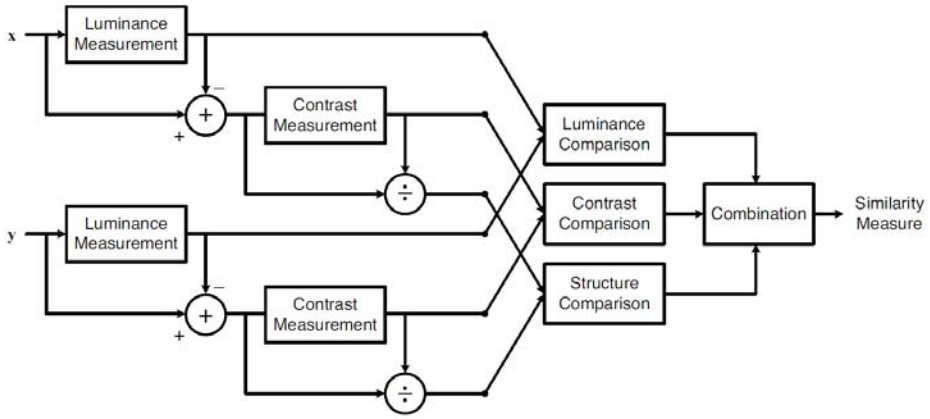


Figure 7.5: *Structure Similarity Index Measurement (SSIM) Block Diagram [2]*

where $l(\mathbf{X}, \mathbf{Y})$ is the luminance comparison function, $c(\mathbf{X}, \mathbf{Y})$ is the contrast comparison function, and $s(\mathbf{X}, \mathbf{Y})$ is the structure comparison function. These three components seem to be the most important aspects for the task of tracking.

7.3.5 Object Detection using Sparse Representation

In this section, we describe a novel object detector based on sparse representation. Our objective is to build a generic object detector using obtained tracking results in previous frames without any prior training. This object detector adaptively learns the object from few frames and then uses the results for detecting the object in the following frames. Therefore, as mentioned in section II, when the object disappears at a certain frame in the sequence, the object detector is invoked to search and locate the object in the following frames. Here, we assume that the object has been tracked in the previous t frames and at frame $t + 1$ it either disappears completely or moves far from the previous location such that the tracker can not follow it. To handle this scenario, we search for the object in frame $t + 1$ using a detector based on sparse

representation and the developed dictionary. This is achieved by sliding a window across the image and deciding whether the window contains the object or not.

There are various definitions of object localization in the literature and they differ in how the location of an object in the image is represented, *e.g.*, center point, contour, a bounding box, or by a pixel-wise segmentation [129], [130]. The goal of our detector is to find a bounding box around the object. Object detection using sliding window rely on evaluating a quality function f, *e.g.*, a classification method, over many rectangular subregions of the image and selecting the window with the maximum score as the object's location follows:

$$R_{obj} = argmax_{R \subseteq image} f(R) \tag{7.10}$$

where R ranges over all rectangular regions in the image. This maximization can not be done exhaustively since it is time consuming. In fact, there are heuristic methods that have been proposed to speed up the search. Most of these methods reduce the number of necessary function evaluations by searching over a coarse gird of possible rectangle locations and by using certain fixed window sizes. In addition, local optimization methods can be used instead of global methods using prior information about background and identifying promising areas where the probability of finding the target is not low. We use Efficient subwindow search method [129] in order to find promising windows in the image. Then, each window is classified either as the tracked object or not using sparse representation with the constructed dictionary as explained in section III. If we find more than one window as candidates for the object, the one with the smallest difference with previous samples is selected and the

algorithm resumes tracking. If the object is not detected in the current frame, we keep looking for it in the following frames.

7.3.6 Feature Extraction

In this section, we describe in detail the features that we use for representing the tracked object as well as the positive and negative samples that are used to build the dictionary.

For feature extraction, we need to address changes that can happen to the appearance of the tracked object (to address challenges in appearance modeling of the target in visual tracking). Scale is one of the main changes that affect the appearance of the object. To address this issue, a multiscale image representation is formed by convolving the input image with a Gaussian filter of different spatial variances [131]. In this representation, the scale space of an image (or sample) is defined as a function, $L(x, y, \sigma)$, that is produced from the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$, with an input image (sample), $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \tag{7.11}$$

The Gaussian filters have to be truncated in the experiments and replaced by rectangle filters. Using rectangle filters can significantly improve the speed of the algorithm without affecting the performance [125]. Therefore, for each sample, its multiscale representation is constructed by convolving the sample with a set of rectangle filters

at multiple scales$F_{1,1}, ..., F_{w,h}$ as:

$$F_{w,h}(x,y) = \begin{cases} \dfrac{1}{wh} & 1 \leq x \leq w, 1 \leq y \leq h \\ \\ 0 & \text{otherwise} \end{cases} \tag{7.12}$$

where w and h are the width and height of the rectangle filter. After this step, each filtered image is used for extracting Haar-like features. The extracted Haar-like feature vector for each sample has a high dimension which leads to a heavy computation burden, and tracking can not be performed in real-time. In addition, a lower dimension feature vector can be discriminative enough for tracking applications as there are only two classes, *i.e.,* object and background. To lower the computational complexity, compressive sensing is used to reduce the dimensionality of the extracted feature vector. Assume, a matrix $\Phi \in \mathbf{R}^{m \times n}(m \ll n)$ satisfies the restricted isometry property of order $k < m$ and isometry constant $0 \leq \delta < 1$ if for all k-sparse signal $\mathbf{u} \in \mathbf{S}_k = \{\alpha \in \mathbf{R}^n : \| \alpha \|_0 = k\}$ [117]:

$$(1-\delta) \parallel \mathbf{u} \parallel_2^2 \leq \parallel \Phi\mathbf{u} \parallel_2^2 \leq (1+\delta) \parallel \mathbf{u} \parallel_2^2 \tag{7.13}$$

In the theory of compressive sensing, if the matrix Φ follows the restricted isometry property, a sparse signal or extracted feature vector, v, can be exactly recovered with an overwhelming probability by using few elements:

$$\mathbf{z} = \Phi\mathbf{v} \tag{7.14}$$

where $\mathbf{z}^{m \times 1}$ is the reduced dimension feature vector.

The compressive sensing theory demonstrates that for a sufficiently high dimension feature space, the features can be projected to a randomly chosen low dimensional

space which contains enough information to reconstruct the original high-dimensional features. This feature reduction method is data-independent and information-preserving.

7.4 Experiments

7.4.1 Tracking Experiments

In this section, we present our experiments for evaluating the proposed approach and present the results on various publicly available video sequences and our own collected video clips. The challenging factors include large pose variation, full and partial occlusion, large scale change, scene blur, significant lighting condition variations and disappearance. The proposed method is implemented in MATLAB (on machine with 3GB) and compared with the well known state-of-the-art tracking systems. In all the experiments, the ground truth center of the object is labled every five frames and the location is interpolated for the other frames. In frames where the object does not show up, the location is labeled NA. The similarity (SSIM) threshold is set to be 0.5. In SL0 algorithm, parameter η is set to 0.1. To balance between the computational complexity and efficiency of ROI modeling, the number of positive and negative samples are assumed to be 50. The initial location of the target is given in all the experiments. We compared our tracker with eight state-of-the-art trackers: fast compressive tracking (FCT) [132], compressive tracking (CT) [133] compressive sensing (CS) tracker [134], Frag [135], OAB [136], MIL tracker [98], multi-task tracker (MTT) [137], and ASLA [138].

7.4.2 Quantitative Evaluation

Nine algorithms including our proposed algorithm were compared using the twelve video sequences. Two evaluation criteria were used to quantitatively assess the performance of the tracking algorithms. The first evaluation criteria is the tracking success rate which is defined in [139] as:

$$score = \frac{area(ROI_T \cap ROI_G)}{area(ROI_T \cup ROI_G)} \tag{7.15}$$

where ROI_T is the tracking bounding box and ROI_G is the ground truth bounding box. The score is evaluated for each frame and if the value is larger than 0.5, the tracking result is acceptable. The success rate for the different tracking algorithms on several video sequences are presented in Table 7.1.

The second criteria is the center location error (CLE) which measures the Euclidean distance between the center of the tracking bounding box and the center of the ground truth bounding box. The results of CLE are summarized in Table 7.2. Quantitative evaluations show that the proposed method compares favourably against the state-of-the-art methods. The proposed method has excellent performance with 87% success rate on the *Ball* sequence, where the ball temporarily leaves the scene and returns after a few frames. The second best method is FCT with only 21% success rate.

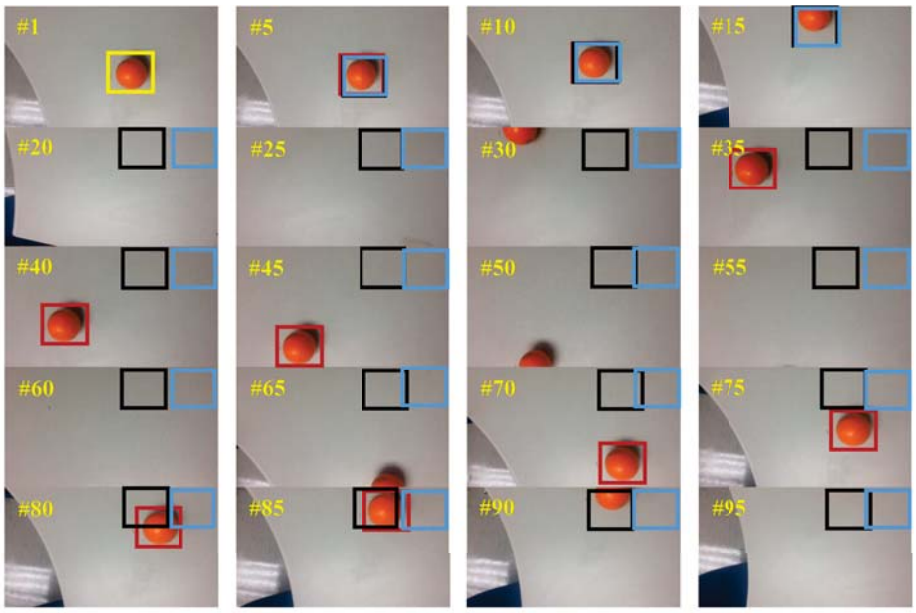


Figure 7.6: *Tracking Results of The Ball Sequence; red: our proposed method, blue: FCT, black: OAB*

7.4.3 Qualitative Evaluation

7.4.3.1 Disappearance

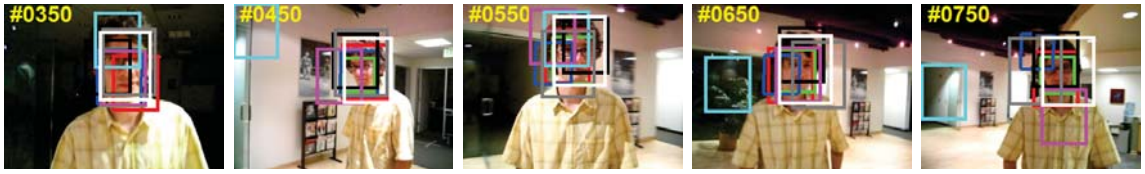
the proposed tracking algorithm is evaluated on a simple video clip which was recorded by the authors. In this video clip, a ball moves on a surface and then moves out of the field of view of the camera for few frames and returns back into the field of view. The exiting and returning locations for the ball can be the same or different. The results on this video clip simply show that most of the tracking algorithms fail to track this simple object under the described scenario when the object temporarily disappears or becomes fully occluded and then reappears. Few sample frames from this video clip are shown in Fig. 7.6. In this sequence, the ball in the middle of the first frame is selected manually. The tracking results for only three tracking systems

are shown in Figure 7.6, where the red colored squares show the results of our tracking system, the blue colored squares show the results of FCT and the black ones show the results of OAB. The success rate (SR) and center location error (CLE) results for all the tracking systems are shown in Table I and Table II. In frames #5 and #10, the ball is moving up and the three trackers work fine and they track the ball accurately. In frame #15, the ball is leaving the scene and half of the ball appears in the frame. Our tracker stops tracking and does not show any results, *i.e.,* there is no red box in the frame. However, all the other tracking systems continue tracking and the results are false tracking in frames #20 and #25. In addition, they can not find the ball after it reappears and they lose track of the target in the following frames (#35, #40, #45). On the other hand, our tracking method detects the ball in frame #35 and continues tracking correctly in frames #40 and #45. Again, in frame #50 our method stops tracking and there are no red boxes in those frames. We note that the locations where the ball leaves in frame #50 and returns in frame #65 are different, which shows that our method is robust and the detection is not dependent on the location. Overall, our algorithm outperforms the state-of-the-art methods on this synthetic video sequence.

7.4.3.2 Illumination and pose changes

few sampled tracking results, of different tracking algorithms, on several video sequences are shown in Fig. 7.7. In the *David sequence*, the person walks from a dark region to a bright region while changing his head pose. The tracking results for the David sequence show that our tracker can handle well both illumination and pose variations as the appearance changes gradually when the person walks out of the

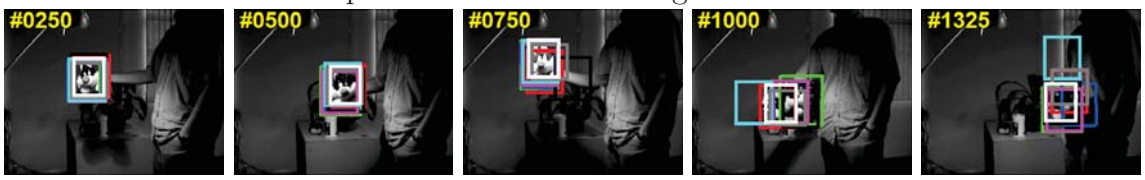
Tracking Results of the *Dark Car* Sequence which has difficult illumination conditions and background clutters



Tracking Results of the *David* Sequence which has large illumination variation, partial occlusion and pose change

Tracking Results of the *Faceocc2* Sequence which has significant and long duration occlusion and in plane rotation

Tracking Results of the *Football* Sequence which has occlusion, in plane rotation, out of plane rotation and background clutters



Tracking Results of the *Sylvester* Sequence which pose change, fast motion and illumination change

Tracking Results of the *Tiger* Sequence which has non-rigid object deformation, motion blur, fast motion and in plane rotation

Figure 7.7: Tracking Results on sample frames from Dark Car, David, FaceOcc2, Football, Sylvester and Tiger video sequences.

Table 7.1: Success Rate (%)

Sequence	FCT	CT	CS	Frag	OAB	MIL	MTT	ASLA	Proposed Method
Ball	21	18	15	17	20	16	18	15	87
Biker	35	84	5	3	66	1	9	10	75
Tiger	52	50	62	19	24	34	24	14	68
Chasing	79	47	67	21	71	65	96	63	81
David	98	94	8	8	32	71	41	34	97
Sylvester	77	69	57	34	65	77	67	82	79
Shaking II	88	55	12	34	74	41	93	82	84
Panda	84	90	1	9	83	80	11	71	75
Football	76	74	35	26	31	77	67	7	69
Dark Car	36	53	6	0	14	48	59	57	55
FaceOcc2	99	99	39	54	49	97	88	93	98
Goat	77	26	26	14	46	27	39	37	71
Average SR	68.5	63.2	27.7	19.9	47.9	52.8	51	47.0	78.5

dark room. Other trackers such as FCT and CT also preform well on this sequence. However, CS and Frag trackers can not track properly, for example the Frag tracker loses the target from frame 380 and up, which shows that these methods are not robust for pose and illumination changes. In the *Sylvester sequence*, the object has pose and illumination changes. ASLA, MIL and our method perform well on most frames of this sequence. However, MTT can not track the target accurately because of using holistic features that are less effective for large pose variations.

7.4.3.3 Rotation and abrupt motion

in the *Shaking sequence*, the person moves his head unpredictably and the head undergoes large appearance variation because of drastic illumination change and unpredictable motion. The FCT method has the best tracking results and our method shows good performance on this sequence because of the use of the detector. When-

Table 7.2: Center Location Error (CLE)

Sequence	FCT	CT	CS	Frag	OAB	MIL	MTT	ASLA	Proposed Method
Ball	45	48	55	52	48	52	50	58	18
Biker	12	6	176	107	10	44	68	109	6
Tiger	23	25	48	39	42	27	61	49	8
Chasing	10	12	9	56	9	13	4	47	8
David	11	14	72	73	57	19	125	57	13
Sylvester	9	14	84	47	12	9	18	9	15
Shaking II	15	46	255	119	18	58	16	27	18
Panda	6	10	157	56	8	7	47	9	10
Football	13	14	43	144	37	13	9	207	24
Dark Car	9	10	89	116	11	9	7	8	8
FaceOcc2	12	16	29	57	36	17	19	20	13
Goat	18	103	137	140	71	109	99	95	25
Average CLE	15.2	26.5	96.1	83.8	29.9	31.4	43.5	57.9	13.1

ever the head movement is large, our method uses the detector and searches for the face in the frame.

One of the challenging sequences is the *Tiger sequence*, where the target experiences changes in appearance such as out of plane rotation with occlusion. Our method performs very well on this sequence since in some frames the target is mostly occluded and other trackers can not find it after reappearing. In frame #346, our method does not show any results as the tiger is mostly occluded but after reappearing in the following frames, the tiger was detected and tracking resumes.

7.4.3.4 Occlusion

In the *goat sequence* there are pose variations, partial occlusion and shape deformation. Most of the trackers can not handle all these changes and the success rate is

less than 50% for CT, CS and Frag. However, FCT has the best performance with 77% success rate and our method is the second with 71% success rate.

In the *Panda sequence*, there are in-plane rotation and partial occlusion. Most of the tracking methods drift after the object undergoes large rotation in frame #103, but FCT, ASLSA and our tracking method perform well on this sequence.

To summarize, the average success rate and center location error are calculated at the end of Table 7.1 and 7.2. Our method obtained 79% success rate which is far better than the other trackers. In addition, the CLE for our method is the lowest and outperforms the other methods.

7.5 Conclusion

In this paper, we presented a robust tracking algorithm using adaptive sparse representation and feedback. The adaptive sparse representation allows for modeling the appearance of the tracked object and handling occlusion and noise. For further robustness, after finding the newly tracked ROI , the similarity measure, SSIM, between the previously tracked ROIs and the newly tracked ROI is measured and if the SSIM is above a threshold, the newly tracked region is similar to the previously tracked ROIs corresponding to the object in the sequence of frames and is used to update the dictionary. On the other hand, if the SSIM is less than the threshold, the newly tracked region is not accepted and the algorithm stops tracking and starts detection using the dictionary. The algorithm does not update the dictionary in this case. Measuring SSIM between the newly tracked ROI and the previous ones is inspired by feedback control systems and leads to significant improvement in the efficiency

of the tracking algorithm. We compared the performance of our method with the state-of-the-art trackers on publicly available video sequences. The obtained results show the accuracy and robustness of our tracking algorithm especially in cases when the tracked object becomes fully occluded or temporarily disappears from the scene. In these cases, our proposed algorithm stops tracking till the object reappears and then resumes tracking. The obtained results using the proposed method outperform the results obtained with the state-of-the-art methods.

CHAPTER 8

Conclusion and Future Work

In this dissertation, we presented sparse representation, l^0 and l^1 norm minimization. In addition, smoothed l^0 norm is used as new algorithm to find the sparsest solution for a system of linear equations. Then we used sparse representation and smoothed l^0 for recognition and classification purposes in the experiments.

In chapter 3, we presented a fully automated system for ear recognition and gender classification using sparse representation. The proposed method was evaluated publicly available data sets, UND collection J ear data set and WVU data set. The obtained results using the SL0 algorithm for classification is far better than other classifiers such as SRC, NN or NS. To improve the proposed algorithm, we plan to fuse facial and ear features for purpose of gender classification in future work.

In chapter 4, a new method for gender classification was presented from facial images using sparse representation. Basis pursuit method was used to formulate the problem in order to find the sparsest solution. The experiments were conducted on the FERET data set and extracting features using Gabor Wavelets. We compared the obtained results from the proposed method with previous methods that used the same data set and our results are more robust that other methods.

In chapter 5, we presented a method for classification based on weighted sparse representation using smoothed l^0 norm with non-negative coefficients. In previous methods, there was no constraint on the coefficients and all the atoms in the dictionary are treated uniformly. While the SRC algorithm does not differentiate between the atoms in the dictionary, our algorithm uses mutual information to find the similarity between the query sample and samples in the gallery, which provides useful weights for more accurate classification. As a result, the weights narrow the search space for the algorithm and help the algorithm to converge faster without being trapped in local extremums. In addition, by adding the non-negative constraint, components of this coefficients vector indicate the contribution of the gallery samples towards the representation of a given query sample. The proposed algorithm was evaluated on two publicly available data sets, the Extended Yale B data set for face recognition and UND data set for ear recognition. Experimental results obtained for our proposed method are not only faster, but also result in a better recognition rate even with a smaller number of training samples. In our future work, we will investigate the use of other similarity measures to enhance the results.

In chapter 6, a robust biometrics recognition system was proposed using joint weighted dictionary learning and smoothed L0 norm. Dictionary learning for sparse representation has shown to improve the results in classification and recognition tasks since class labels can be used in obtaining the atoms of learnt dictionary. We proposed a joint weighted dictionary learning which simultaneously learns from a set of training samples an over complete dictionary along with weight vectors that correspond to the atoms in the learnt dictionary. The components of the weight vector associated with an atom represent the relationship between the atom and each of the classes. The

weight vectors and atoms are jointly obtained during the dictionary learning. In the proposed method, a constraint is imposed on the correlation between the obtained atoms that represent different classes to decrease the similarity between these atoms. In addition, we use smoothed L0 norm which is a fast algorithm to find the sparsest solution. Experiments conducted on the West Virginia University (WVU) and the University of Notre Dame (UND) datasets for ear recognition show that the proposed method outperforms other state-of-the-art classifiers.

In chapter 7, we proposed robust object tracking using adaptive sparse representation and feedback and an automatic object tracking system is developed that can handle pose variations, scale variations and temporary disappearance of the object from the scene. We focus on automatic tracking with no prior knowledge other than the location of the region to be tracked in the first frame, which can either be located manually or using a detector that finds the region of interest (ROI). The visual tracking is a binary classification problem. The positive samples are bounding boxes that have high overlap with current position of the target while negative samples are drawn from regions outside the ROI to model background close to the target. The tracking algorithm uses the dictionary to locate the ROI in the following frames via adaptive sparse representation. One of the main issues in tracking systems is false tracking when the object disappears from the scene. Motivated by the concept of feedback in control systems, we overcome the problem of false tracking when the object disappears by comparing the newly tracked region with previous regions to confirm that the object is still in the frame. A structural similarity measure is used to measure similarity between a newly tracked ROI and the previously tracked ROIs and if the similarity is below a certain threshold, the object is assumed to be out of

the scene. In fact, this similarity evaluation is like a feedback loop in our tracking algorithm which makes our method robust, reliable and accurate when compared to the state-of-the-art methods on challenging sequences. If the object is not located in the current frame, the algorithm stops tracking and starts searching for the object in the following frames. The searching is achieved by using a detector based on sparse representation and an adaptive dictionary to efficiently locate the object when it reappears in the scene. Experiments on video sequences, for both quantitative and qualitative evaluations, demonstrate the effectiveness and robustness of the proposed tracking system.

Bibliography

[1] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *Proceedings of the 11th European conference on Computer vision: Part VI*, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1888212.1888247

[2] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, April 2004.

[3] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse representation for ear biometrics," in *Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II*, 2008.

[4] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289 –1306, 2006.

[5] W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, and Z. Zhu, "Robust face recognition via occlusion dictionary learning," *Pattern Recognition*, vol. 47, no. 4, pp. 1559 – 1572, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320313004469

[6] Y. Zhou, K. Liu, R. E. Carrillo, K. E. Barner, and F. Kiamilev, "Kernel-based sparse representation for gesture recognition," *Pattern Recognition*, vol. 46, no. 12, pp. 3208 – 3222, 2013.

[7] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 12 – 23, 2014.

[8] Y. Xie, W. Zhang, Y. Qu, and Y. Zhang, "Discriminative subspace learning with sparse representation view-based model for robust visual tracking," *Pattern Recognition*, vol. 47, no. 3, pp. 1383 – 1394, 2014.

[9] M. Yang, L. Zhang, S. C. Shiu, and D. Zhang, "Gabor feature based robust representation and classification for face recognition with gabor occlusion dictionary," *Pattern Recognition*, vol. 46, no. 7, pp. 1865 – 1878, 2013.

[10] A. Taalimi, R. Khorsandi, and H. Qi, "Online multi-modal task-driven dictionary learning and robust joint sparse representation for visual tracking," in *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, Aug 2015.

[11] R. Khorsandi, A. Taalimi, M. Abdel-Mottaleb, and H. Qi, "Joint weighted dictionary learning and classifier training for robust biometric recognition," in *Global Conference on Signal and Information Processing (GlobalSIP), 2015 IEEE*, Dec 2015.

[12] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311 –4322, Nov. 2006.

[13] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 27 –35, Jan. 2009.

[14] A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 358 –371, Aug. 2010.

[15] C.-K. Chiang, C.-H. Liu, C.-H. Duan, and S.-H. Lai, "Learning component-level sparse representation for image and video categorization," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4775–4787, Dec 2013.

[16] W. Nunziati, S. Sclaroff, and A. Del Bimbo, "Matching trajectories between video sequences by exploiting a sparse projective invariant representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 517–529, March 2010.

[17] A. Taalimi, S. Ensafi, H. Qi, S. Lu, A. Kassim, and C. L. Tan, "Multimodal dictionary learning and joint sparse representation for hep-2 cell classification," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2015.

[18] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2259–2272, 2011.

[19] J. Xu and J. Yang, "A nonnegative sparse representation based fuzzy similar neighbor classifier," *Neurocomputing*, vol. 99, no. 0, pp. 76 – 86, 2013.

[20] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210 –227, 2009.

[21] R. Khorsandi, A. Taalimi, and M. Abdel-Mottaleb, "Joint weighted dictionary learning and classifier training for robust biometric recognition," in *Biometrics: Theory, Applications and Systems (BTAS), 2015 IEEE Seventh International Conference on*, Sept 2015.

[22] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed norm," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 289 –301, Jan. 2009.

[23] R. Khorsandi, S. Cadavid, and M. Abdel-Mottaleb, "Ear recognition via sparse representation and gabor filters," in *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2012, pp. 278–282.

[24] R. Khorsandi and M. Abdel-Mottaleb, "Gender classification using 2-d ear images and sparse representation," in *IEEE Workshop on Applications of Computer Vision (WACV)*, 2013, pp. 461–466.

[25] D. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 471–478.

[26] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[27] R. Khorsandi and M. Abdel-Mottaleb, "Ear biometrics and sparse representation based on smoothed l0 norm," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 08, p. 1456016, 2014.

[28] ——, "Gender classification using facial images and basis pursuit," in *Computer Analysis of Images and Patterns*. Springer, 2013, pp. 294–301.

[29] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *In Adv. NIPS*, 2006.

[30] S. Gao, I.-H. Tsang, and L.-T. Chia, "Kernel sparse representation for image classification and face recognition," in *Computer Vision ECCV 2010*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, vol. 6314, pp. 1–14.

[31] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang, "Learning with l^1 -graph for image analysis," *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 858–866, 2010.

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

[33] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition under variable lighting and pose," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 954 –965, 2012.

[34] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *17th IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 1601–1604.

[35] C.-Y. Lu, H. Min, J. Gui, L. Zhu, and Y.-K. Lei, "Face recognition via weighted sparse representation," *Journal of Visual Communication and Image Representation*, vol. 24, no. 2, pp. 111 – 116, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S104732031200082X

[36] L. Duarte, R. Ando, R. Attux, Y. Deville, and C. Jutten, "Separation of sparse signals in overdetermined linear-quadratic mixtures," in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds. Springer Berlin Heidelberg, 2012, vol. 7191, pp. 239–246.

[37] R. Khorsandi, M. Abdel-Mottaleb, and H. Qi, "Robust object tracking via adaptive sparse representation," in *Global Conference on Signal and Information Processing (GlobalSIP), 2015 IEEE*, Dec 2015.

[38] A. Taalimi and H. Qi, "Robust multi-object tracking using confident detections and safe tracklets," in *Image Processing (ICIP), 2015 22nd IEEE International Conference on*. IEEE, 2015.

[39] X. Gao, K. Zhang, D. Tao, and X. Li, "Image super-resolution with sparse neighbor embedding," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3194–3205, 2012.

[40] M. Burge and W. Burger, "Ear biometrics," in *BIOMETRICS: Personal Identification in a Networked Society*, 1998.

[41] D. Hurley, M. Nixon, and J. Carter, "Automatic ear recognition by force field transformations," in *IEE Colloquium on Visual Biometrics (Ref.No. 2000/018)*, 2000.

[42] M. Abdel-Mottaleb and J. Zhou, "A system for ear biometrics from face profile images," in *Proc. of the first ICGST International Conference on Graphics, Vision and Image Processing GVIP*, vol. 05, Dec. 2005.

[43] A. Kumar and T.-S. T. Chan, "Robust ear identification using sparse representation of local texture descriptors," *Pattern Recognition*, vol. 46, no. 1, pp. 73 – 85, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312002907

[44] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," *Comm. Pure Appl. Math*, vol. 59, pp. 797–829, 2004.

[45] M. Hagan and M. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989 –993, Nov 1994.

[46] A. Ghaffari, M. Babaie-Zadeh, and C. Jutten, "Sparse decomposition of two dimensional signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[47] D.-Y. Chen and K.-Y. Lin, "Robust gender recognition for real-time surveillance system," in *IEEE International Conference on Multimedia and Expo (ICME)*, July 2010.

[48] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang, "Gender recognition from body," in *Proceedings of the 16th ACM international conference on Multimedia*, 2008. [Online]. Available: http://doi.acm.org/10.1145/1459359.1459470

[49] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, "Sexnet: A neural network identifies sex from human faces," *Proceedings Conf. Advances in Neural Information Processing Systems 3*, pp. 572 –577, 1990.

[50] R. Khorsandi and M. Abdel-Mottaleb, "Gender classification using facial images and basis pursuit," in *15th International Conference on Computer Analysis of Images and Patterns*, 2013.

[51] H. Harb and L. Chen, "Gender identification using a general audio classifier," in *Proceedings of the International Conference on Multimedia and Expo - Volume 1*, 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=1153922.1154344

[52] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, "A study on gait-based gender classification," *IEEE Transactions on Image Processing*, vol. 18, no. 8, pp. 1905 –1910, 2009.

[53] H. Chen and B. Bhanu, "Human ear recognition in 3d," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 718 –737, April 2007.

[54] S. Cadavid, S. Fathy, J. Zhou, and M. Abdel-Mottaleb, "An adaptive resolution voxelization framework for 3d ear recognition," in *International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1 –6.

[55] Z. Huang, Y. Liu, C. Li, M. Yang, and L. Chen, "A robust face and ear based multimodal biometric system using sparse representation," *Pattern Recognition*, vol. 46, no. 8, pp. 2156 – 2168, 2013.

[56] K. Chang, K. Bowyer, S. Sarkar, and B. Victor, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1160 – 1165, Sept. 2003.

[57] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb, "An efficient 3-d ear recognition system employing local and holistic features," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 978 –991, June 2012.

[58] P. Yan and K. Bowyer, "Biometric recognition using 3d ear shape," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1297 –1308, 2007.

[59] B. Moghaddam and M.-H. Yang, "Gender classification with support vector machines," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[60] B. Wu, H. Ai, and C. Huang, "Facial image retrieval based on demographic classification," in *Proceedings of the 17th International Conference on Pattern Recognition, ICPR*, vol. 3, 2004.

[61] G. W. Cottrell and J. Metcalfe, "Empath: face, emotion, and gender recognition using holons," in *Proceedings of the 1990 conference on Advances in neural information processing systems 3*, 1990. [Online]. Available: http://dl.acm.org/citation.cfm?id=118850.105194

[62] S. Gutta and H. Wechsler, "Gender and ethnic classification of human faces using hybrid classifiers," in *International Joint Conference on Neural Networks, IJCNN*, vol. 6, 1999.

[63] P. Gnanasivam and S. Muttan, "Gender classification using ear biometrics," in *Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP 2012)*, ser. Lecture Notes in Electrical Engineering, M. S and S. S. Kumar, Eds. Springer India, 2013, vol. 222, pp. 137–148.

[64] P. Yan and K. Bowyer, "Empirical evaluation of advanced ear biometrics," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, 2005. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.450

[65] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb, "A computationally efficient approach to 3d ear recognition employing local and holistic features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2011.

[66] J. Bustard and M. Nixon, "Toward unconstrained ear recognition from two-dimensional images," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 40, no. 3, pp. 486 –494, 2010.

[67] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467 –476, 2002.

[68] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. Nixon, "A survey on ear biometrics," *ACM Computing Surveys*, 2011. [Online]. Available: http://eprints.soton.ac.uk/272951/

[69] S. S. Chen, D. L. Donoho, Michael, and A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[70] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[71] R.-L. Hsu, M. Abdel-Mottaleb, and A. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696 –706, may 2002.

[72] A. Jain, J. Huang, and S. Fang, "Gender identification using frontal facial images," in *IEEE International Conference on Multimedia and Expo, ICME 2005*, july 2005, p. 4 pp.

[73] H. Lu, Y. Huang, Y. Chen, and D. Yang, "Automatic gender recognition based on pixel-pattern-based texture feature," *Journal of Real-Time Image Processing*, vol. 3, pp. 109–116, 2008. [Online]. Available: http://dx.doi.org/10.1007/s11554-008-0072-2

[74] P. O. Hoyer, "Non-negative sparse coding," in *Neural Networks For Signal Processing Xii (Proc. Ieee Workshop On Neural Networks For Signal Processing)*, 2002, pp. 557–565.

[75] M. Lyons, J. Budynek, A. Plante, and S. Akamatsu, "Classifying facial attributes using a 2-d gabor wavelet representation and discriminant analysis," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[76] M. Abdel-Mottaleb, "Image retrieval based on edge representation," in *International Conference on Image Processing*, vol. 3, 2000, pp. 734–737 vol.3.

[77] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893 vol. 1.

[78] O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes, "Trainable classifier-fusion schemes: An application to pedestrian detection," in *12th International IEEE Conference on Intelligent Transportation Systems*, Oct 2009, pp. 1–6.

[79] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[80] R. Baraniuk, E. Candes, M. Elad, and Y. Ma, "Applications of sparse representation and compressive sensing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 906 –909, 2010.

[81] Z. Jiang, Z. Lin, and L. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 11, pp. 2651–2664, Nov 2013.

[82] K. Skretting and K. Engan, "Recursive least squares dictionary learning algorithm," *Signal Processing, IEEE Transactions on*, vol. 58, no. 4, pp. 2121–2130, April 2010.

[83] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 791–804, April 2012.

[84] J. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *Image Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1395–1408, July 2009.

[85] D. Bradley and J. A. D. Bagnell, "Differentiable sparse coding," in *Proceedings of Neural Information Processing Systems 22*, December 2008.

[86] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 791–804, April 2012.

[87] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted 1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[88] S. Guo, Q. Ruan, and Z. Miao, "Similarity weighted sparse representation for classification," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, Nov 2012, pp. 1241–1244.

[89] R. Sznitman, R. Richa, R. Taylor, B. Jedynak, and G. Hager, "Unified detection and tracking of instruments during retinal microsurgery," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 5, pp. 1263–1273, May 2013.

[90] J. Martinez-del Rincon, M. Lewandowski, J.-C. Nebel, and D. Makris, "Generalized laplacian eigenmaps for modeling and tracking human motions," *Cybernetics, IEEE Transactions on*, vol. 44, no. 9, pp. 1646–1660, Sept 2014.

[91] E.-J. Ong and R. Bowden, "Robust facial feature tracking using shape-constrained multiresolution-selected linear predictors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1844–1859, Sept 2011.

[92] A. Jepson, D. Fleet, and T. El-Maraghi, "Robust online appearance models for visual tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 10, pp. 1296–1311, Oct 2003.

[93] X. Yu, J. Yang, T. Wang, and T. Huang, "Key point detection by max pooling for tracking," *Cybernetics, IEEE Transactions on*, vol. 45, no. 3, pp. 444–452, March 2015.

[94] H. Shen, S. Li, J. Zhang, and H. Chang, "Tracking-based moving object detection," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 3093–3097.

[95] W. Ni and A. Caplier, "Adaptive appearance face tracking with alignment feedbacks," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, Sept 2012, pp. 1825–1828.

[96] R. Mosberger and H. Andreasson, "An inexpensive monocular vision system for tracking humans in industrial environments," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, May 2013, pp. 5850–5857.

[97] V. Kyrki and D. Kragic, "Tracking rigid objects using integration of model-based and model-free cues," *Machine Vision and Applications*, vol. 22, no. 2, pp. 323–335, 2011. [Online]. Available: http://dx.doi.org/10.1007/s00138-009-0214-y

[98] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1619–1632, Aug 2011.

[99] Z. Ji and W. Wang, "Robust object tracking via multi-task dynamic sparse model," in *Image Processing (ICIP), 2014 IEEE International Conference on*, Oct 2014, pp. 393–397.

[100] F. Azhar and T. Tjahjadi, "Significant body point labeling and tracking," *Cybernetics, IEEE Transactions on*, vol. 44, no. 9, pp. 1673–1685, Sept 2014.

[101] Z. Zhou, Y. Wang, and E. K. Teoh, "Robust object tracking using bi-model," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 3103–3107.

[102] T. Bai, Y.-F. Li, and X. Zhou, "Learning local appearances with sparse representation for robust and fast visual tracking," *Cybernetics, IEEE Transactions on*, vol. 45, no. 4, pp. 663–675, April 2015.

[103] Y. Xie, W. Zhang, C. Li, S. Lin, Y. Qu, and Y. Zhang, "Discriminative object tracking via sparse representation and online dictionary learning," *Cybernetics, IEEE Transactions on*, vol. 44, no. 4, pp. 539–553, April 2014.

[104] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2259–2272, Nov 2011.

[105] F. Chen, Q. Wang, S. Wang, W. Zhang, and W. Xu, "Object tracking via appearance modeling and sparse representation," *Image and Vision Computing*, vol. 29, no. 11, pp. 787 – 796, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885611000825

[106] J. Wang and Y. Yagi, "Many-to-many superpixel matching for robust tracking," *Cybernetics, IEEE Transactions on*, vol. 44, no. 7, pp. 1237–1248, July 2014.

[107] C.-H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 685–692.

[108] T. Yang, B. Li, and M.-H. Meng, "Robust object tracking with reacquisition ability using online learned detector," *Cybernetics, IEEE Transactions on*, vol. 44, no. 11, pp. 2134–2142, Nov 2014.

[109] Y. Wu, M. Pei, M. Yang, J. Yuan, and Y. Jia, "Robust discriminative tracking via landmark-based label propagation," *Image Processing, IEEE Transactions on*, vol. 24, no. 5, pp. 1510–1523, May 2015.

[110] M. Black and A. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63–84, 1998.

[111] R. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1631–1643, Oct 2005.

[112] W. Hu, X. Zhou, W. Li, W. Luo, X. Zhang, and S. Maybank, "Active contour-based visual tracking by integrating colors, shapes, and motions," *Image Processing, IEEE Transactions on*, vol. 22, no. 5, pp. 1778–1792, May 2013.

[113] C.-P. Wei, C.-F. Chen, and Y.-C. Wang, "Robust face recognition with structurally incoherent low-rank matrix decomposition," *Image Processing, IEEE Transactions on*, vol. 23, no. 8, pp. 3294–3307, Aug 2014.

[114] Z. Tian and G. Giannakis, "Compressed sensing for wideband cognitive radios," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–1357–IV–1360.

[115] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," DTIC Document, Tech. Rep., 2008.

[116] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[117] E. Candes and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, Dec 2005.

[118] A. Moghadam, M. Kumar, and H. Radha, "Common and innovative visuals: A sparsity modeling framework for video," *Image Processing, IEEE Transactions on*, vol. 23, no. 9, pp. 4055–4069, Sept 2014.

[119] M. Azab, H. Shedeed, and A. Hussein, "New technique for online object tracking-by-detection in video," *Image Processing, IET*, vol. 8, no. 12, pp. 794–803, 2014.

[120] G. Cui, J. Wang, and J. Li, "Robust multilane detection and tracking in urban scenarios based on lidar and mono-vision," *Image Processing, IET*, vol. 8, no. 5, pp. 269–279, May 2014.

[121] F. Pernici and A. Del Bimbo, "Object tracking by oversampling local features," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 12, pp. 2538–2551, Dec 2014.

[122] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *Image Processing, IEEE Transactions on*, vol. 22, no. 7, pp. 2559–2573, July 2013.

[123] L. Vacchetti, V. Lepetit, and P. Fua, "Stable real-time 3d tracking using online and offline information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 10, pp. 1385–1391, Oct 2004.

[124] S. Taylor and T. Drummond, "Multiple target localisation at over 100 fps," 2009.

[125] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[126] M. Sampat, Z. Wang, S. Gupta, A. Bovik, and M. Markey, "Complex wavelet structural similarity: A new image similarity index," *Image Processing, IEEE Transactions on*, vol. 18, no. 11, pp. 2385–2401, Nov 2009.

[127] R. Vieux, J. Benois-Pineau, J. Domenger, and A. Braquelaire, "Espi image indexing and similarity search in radon transform domain," in *Content-Based Multimedia Indexing, 2009. CBMI '09. Seventh International Workshop on*, June 2009, pp. 231–236.

[128] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu, "Robust face recognition via adaptive sparse representation," *Cybernetics, IEEE Transactions on*, vol. 44, no. 12, pp. 2368–2378, Dec 2014.

[129] C. H. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.

[130] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 606–613.

[131] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[132] K. Zhang, L. Zhang, and M.-H. Yang, "Fast compressive tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 10, pp. 2002–2015, Oct 2014.

[133] ——, "Real-time compressive tracking," in *Computer Vision ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, vol. 7574, pp. 864–877.

[134] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 1305–1312.

[135] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, June 2006, pp. 798–805.

[136] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2006, pp. 6.1–6.10, doi:10.5244/C.20.6.

[137] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 2042–2049.

[138] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 1822–1829.

[139] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.