A GRAPH BASED APPROACH FOR FINDING PEOPLE IN NEWS

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Derya Özkan

July, 2007

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Pınar Duygulu Şahin(Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Selim Aksoy

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Fatoş Yarman Vural

Approved for the Institute of Engineering and Science:

Prof. Dr. Mehmet B. Baray
Director of the Institute

ABSTRACT

A GRAPH BASED APPROACH FOR FINDING PEOPLE IN NEWS

Derya Özkan
M.S. in Computer Engineering
Supervisor: Asst. Prof. Dr. Pınar Duygulu Şahin
July, 2007

Along with the recent advances in technology, large quantities of multi-modal data has arisen and became prevalent. Hence, effective and efficient retrieval, organization and analysis of such data constitutes a big challenge. Both news photographs on the web and news videos on television form this kind of data by covering rich sources of information. People are mostly the main subject of the news; therefore, queries related to a specific person are often desired.

In this study, we propose a graph based method to improve the performance of person queries in large news video and photograph collections. We exploit the multi-modal structure of the data by associating text and face information. On the assumption that a person's face is likely to appear when his/her name is mentioned in the news, only the faces associated with the query name are selected first to limit the search space for a query name. Then, we construct a similarity graph of the faces in this limited search space, where nodes correspond to the faces and edges correspond to the similarity between the faces. Among these faces, there could be many faces corresponding to the queried person in different conditions, poses and times. There could also be other faces corresponding to other people in the news or some non-face images due to the errors in the face detection method used. However, in most cases, the number of corresponding faces of the queried person will be large, and these faces will be more similar to each other than to others. To this end, the problem is transformed into a graph problem, in which we seek to find the densest component of the graph. This most similar subset (densest component) is likely to correspond to the faces of the query name. Finally, the result of the graph algorithm is used as a model for further recognition when new faces are encountered. In the paper, it has been

shown that the graph approach can also be used for detecting the faces of the anchorpersons without any supervision.

The experiments are performed on two different data sets: news photographs and news videos. The first set consists of thousands of news photographs from Yahoo! news web site. The second set includes 229 broadcast news videos provided by NIST for TRECVID 2004. Images from the both sets are taken in real life conditions and, therefore, have a large variety of poses, illuminations and expressions. The results show that proposed method outperforms the text only based methods and provides cues for recognition of faces on the large scale.

Keywords: Face recognition, face retrieval, SIFT features.

ÖZET

HABERLERDE KİŞİLERİ BULMAYA YARAYAN ÇİZGEYE DAYALI BİR YÖNTEM

Derya Özkan
Bilgisayar Mühendisliği,, Yüksek Lisans
Tez Yöneticisi: Asst. Prof. Dr. Pıinar Duygulu Şahin
Temmuz, 2007

Gelişen teknoloji ile birlikte geniş ve çok-modelli veri kümeleri yaygın hale gelmiştir. Bu veri kümelerinin etkin ve hızlı bir şekilde erişimi, düzenlenmesi ve analizi büyük bir ilgi alanı oluşturmaktadır. Hem internet üzerindeki haber resimleri ve hem de televizyondaki haber görüntüleri kendi içinde bir çok bilgiyi barındıran önemli veri kaynaklarıdır. Kişiler genellikle haberin ana konusu olup, bu kişileri sorgulama önemli ve çoğu zaman istenen bir işlemdir.

Bu çalışmada, haber fotoğrafları ve görüntülerinden oluşan geniş veri kümelerinde kişilerin sorgulanmasını sağlayan çizgeye dayalı bir yöntem sunulmuştur. Yöntem isim ve yüzlerin ilişkilendirilmesine dayanmaktadır. Haber başlığında kişinin ismi geçiyor ise fotoğrafta da o kişinin yüzünün bulunacağı varsayımıyla, ilk olarak sorgulanan isim ile ilişkilendirilmiş, fotoğraflardaki tüm yüzler seçilir. Bu yüzler arasında sorgu kişisine ait farklı koşul, poz ve zamanlarda çekilmiş, pek çok resmin yanında, haberde ismi geçen başka kişilere ait yüzler ya da kullanılan yüz bulma yönteminin hatasından kaynaklanan yüz olmayan resimler de bulunabilir. Yine de, çoğu zaman, sorgu kişisine ait resimler daha çok olup, bu resimler birbirine diğerlerine olduğundan daha çok benzeyeceklerdir. Bu nedenle, yüzler arasındaki benzerlikler çizgesel olarak betimlendiğinde, birbirine en çok benzeyen yüzler bu çizgede en yoğun bileşen olacaktır. Bu çalışmada, sorgu ismiyle ilişkilendirilmiş, yüzler arasında birbirine en çok benzeyen alt kümeyi bulan, çizgeye dayalı bir yöntem sunulmaktadır. çizgeye bağlı yaklaşımla bulunan bu sonuç, daha sonra yeni karşılaşılan yüzlerin tanınmasında da model olarak kullanılabilmektedir. Aynı zamanda, çalışmada sunulan çizgeye dayalı yaklaşım haber görüntülerindeki spikerlerin otomatik olarak bulunması ve elenmesinde de kullanılmıştır.

Deneyler iki ayrı veri kümesi kullanılarak gerçekleştirilmiştir: haber

fotoğrafları ve haber görüntüleri. İlk veri kümesi Yahoo! Haber kanalı üzerinden toplanmiş binlerce resimden oluşmaktadır. Ikinci küme ise NIST tarafından TRECVID 2004 yarışması için sağlanan 229 haber görüntüsünden oluşmaktadir. Resimler gercek hayattan alınmış olduğundan yüzler poz, ışıklandırma ve ifade olarak çok fazla çeşitlilik göstermektedir. Deneylerde elde edilen sonuçlar, sadece isim bazlı sonuçlara göre daha iyi olup, büyük çapta yüz tanıma için fikir vermektedir.

Anahtar sözcükler: Yüz tanıma, yüz sorgulama, SIFT gösterimi.

Acknowledgement

I would first like to express my gratitude to my advisor Dr. Pinar Duygulu Şahin for her guidance and support throughout my studies. I learned a lot from her. Besides her valuable comments and teaching, she was my main source of morale support by highly motivating me evermore. I am also honored to be her first masters student.

I am gratefully thankful to Dr. Selim Aksoy for his suggestions and valuable comments. He always helped me with his deep knowledge whenever I asked for advice.

I would like to thank to my thesis committee members, Dr. Selim Aksoy and Prof. Fatoş Yarman Vural, for reviewing my thesis and making constructive remarks.

I am very thankful to all my friends and Retina team members for all their morale support. It was very much pleasure for me to carry on my studies in such a nice environment.

And finally, I would like to very much thank to my family for their encouragement and invaluable supports.

This research is partially supported by TÜBİTAK Career grant number 104E065 and grant number 104E077.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Along with the recent advances in technology, large quantities of multi-modal data has become more available and widespread. With its emergence, effective and efficient retrieval, organization and analysis of such kind of data has become a challenging problem and aroused interest. News photographs on the web (Figure 1.1) and news videos on television (Figure 1.2) are two examples of this type of data. They acquire rich sources of information wherein. Hence, accessing them is especially important. This importance has also been acknowledged by NIST in TRECVID video retrieval evaluation by choosing the news videos as the data source [1].

People are usually the main subject of the stories in the news. Therefore, queries related to a specific person is often a desired task. In general, a person visually appears when his/her name is mentioned in the news. On this account, the common approach to retrieve information related to a person is to search using his/her name in the associated caption of the news photographs or in the associated speech transcript of the video shots. However, such an approach is likely to yield incorrect results since the retrieved photos/shots associated with the name may not include the query person or any people at all (for instance US

Figure 1.1: Sample news photographs and their associated captions.

president's name may be mentioned while the White House is on the screen, see Figure 1.1).

Detecting faces and eliminating the photos/shots which do not include any face can handle the second problem within the limitations of the selected face detection method. Besides the query person, there might be other people in the story that also appear in the same photo/shot. Hence, a more difficult problem arises in this case, since multiple faces are associated with multiple names and it is not known which face goes with which face. Figure 1.3 exemplifies some of the detected faces in news photographs that are associated with the query name *President George W. Bush*. Even if there are the faces of the query person, there are also the faces of other people and some non-face images that are associated with the same name.

In news videos the problem become more challenging due to the time shift, which usually arises between the appearance of the name and the visual appearance of the person. For example a person's name is mentioned while the anchor

Figure 1.2: Sample shots from news videos and the speech transcript texts aligned with those shots.

Figure 1.3: Sample detected faces that are associated with the name *President George W. Bush*.

person is introducing the related story, but the person actually appears later in the video (Figure 1.4). Therefore, retrieving the faces in the shot which is temporally aligned with the speech transcript including the name usually yields incorrect results and mostly returns faces of the anchorperson/reporter.

The solution to the mismatch between the faces and the names requires the incorporation of a face recognition algorithm. On the other hand, face recognition is a long standing and a difficult problem (see [22, 54, 23, 49, 47] for recent surveys). Although many different approaches have been proposed for recognizing faces, most of the current face recognition methods are evaluated only in controlled environments and for limited datasets. However, for larger and more realistic datasets like news photographs and/or videos, face recognition is still

Figure 1.4: Key-frames from two different videos. The numbers below each image show the distance to shot, in which the name 'Clinton' is mentioned. Note that in both cases, Clinton does not appear visually in the shot in which his name is mentioned but appears in preceding (up image) or succeeding shots (bottom image).

difficult and error-prone due to the noisy and complicated nature of these sets. These sets contain large variations in pose, illumination and facial expression, which cause the face recognition problem even more difficult.

It has been shown in recent studies [44, 7, 8] that the face recognition problem can be simplified by transforming it into a face-name association problem. In this direction, we propose a method for improving the performance of person queries in news datasets by exploiting from both text and visual information.

Our method is built on the observation that when the faces in the photos/shots associated with a given query name is considered, the faces of the query person are likely to be the most frequently appearing ones than any other person, although there may be faces corresponding to other people in the story, or some non-face images due to the face detection algorithm used. Even if the expressions or poses vary, different appearances of the face of the same person tend to be more similar to each other than to the faces of others. In other words, among the group of faces associated with the name, faces of the query person are the ones which are most similar to each other and therefore forms the largest group of similar faces.

In this context, we transform the face-name association problem into the problem of finding the largest group of most similar faces from a set of faces associated with the query name. We approach the problem as a problem of finding the densest component in a graph, which represents the similarities between faces.

To this end, we first find the faces associated with the query name to limit our search space. Then, the similarity of the faces are represented in a graph structure, where the nodes correspond to faces and the edges correspond to face similarities. The problem is then turned into finding the densest component of the graph. This densest component refers to the largest set of highly connected nodes in the graph; thus the largest group of similar faces corresponding to faces of the query person. When we find the faces of the query person with the proposed approach, the returned solution is also used as a model for recognizing new faces.

Different from the most of current face recognition systems, we find the similarity of the two faces based on interest points extracted from those faces. The proposed method benefits from the scale and illumination invariance characteristics of interest features. Besides, it is less sensitive to variations in noise, occlusion, and illumination; and works also in the absence of any of the facial features.

The proposed method is not a solution to the general face recognition problem. Rather, it is a method to increase the retrieval performance of the person queries in the large data sets where names and faces appear together and where traditional face recognition systems cannot be used. It does not require a training step for a specific person and therefore, there is no limitation on the number of people queried.

In the following two sections, we briefly describe the overall approach, and then present the organization of thesis.

1.2 Summary of Contributions

Our person finding approach is based on limiting the search space of a query person first by using text information and then solving the problem by transforming it to a graph problem. After finding the faces of the query person, we use the result as a model for recognizing new faces. The overall approach consists of four steps: constructing a limited search space for a query person by using text and name information, defining similarities between faces in this search space to form a similarity graph of faces, finding the densest component of this graph which corresponds to the faces of the query person, and finally using the result as a model further in recognizing new faces. The main steps of the proposed method are given in Figure 1.6.

In the first step, we use the text information to limit our search space for a query name. It consists of searching for the query name in the caption or in the speech transcript, and choosing the photos/shots that are associated with the query name. As stated earlier, there is usually a time shift between the appearance of a name and the appearance of the face belonging to that name in news videos. This problem is handled by taking a window around the name. The solution is, rather than searching the faces only on the shots including the name of the person, also to include the preceding and succeeding shots.

In the second step, we assign a similarity measure to each pair of faces in the limited search space and represent these similarities among faces in a graph structure. In this similarity graph, nodes correspond to the faces and the edges correspond to the similarities between faces.

In this study, the similarities between faces are represented by using interest points extracted from the faces. We use Lowe's SIFT features, which have been shown to be successful in recognizing objects [34, 38, 25] and faces [26]. Different from the original matching metric of SIFT, we assign the interest points having the minimum distance as the initial matching points. Then, we apply two constraint on these matched points: geometrical constraint and unique match constraint, to eliminate the false matches and select the best matching points.

Finally, the average distance of matching points between two faces is used to assign the distance between these two faces. Using the SIFT features, we both exploit from the scale and illumination invariant property of these features and are able to assign a distance metric between two faces even in the absence of any of facial features.

In the constructed similarity graph, the nodes corresponding to the faces of the query person will be more strongly connected to each other than to other nodes corresponding to other faces in the graph. Moreover, the query person is the one whose face usually appears the most frequently in its search space. Considering all these, the problem is transformed into a graph problem in the third step and solved by a greedy graph algorithm that returns the densest component of the graph. This densest component refers to the set of highly connected nodes in the graph; thus the set of most similar faces corresponding to the faces of the query person.

The final step involves using the result of the graph algorithm as a model to recognize new faces. With the proposed graph algorithm, we can automatically obtain the labeled training set for learning the model in recognizing new faces. We propose two different techniques to use the result based on: degree, distance, and match number modeling. In all those techniques, the model is trained by using the returned graph as the training set.

The method is applied on two different news datasets: news photographs and news videos. The first dataset, namely the news photographs collected by Berg *et al.* [7], is quite different from most of the existing datasets (see Figure 1.5). It consists of large number of photographs with associated captions collected from Yahoo! News on the Web. Photographs are taken in real life conditions rather than in restricted and controlled environments. Therefore, they represent a large variety of poses, illuminations and expressions. They are taken both indoors and outdoors. The large variety of environmental conditions, occlusions, clothing, and ages make the dataset even more difficult to be recognized.

The second dataset is broadcast news videos provided by NIST for TRECVID

Figure 1.5: Sample faces from news photographs [7] (on the top), and news videos (on the bottom).

2004 video retrieval evaluation [1]. Due to the higher noise level and lower resolution, news videos is a harder dataset to work with. In order to handle the problem due to anchorperson faces, we add a mechanism to detect and remove the anchorpeople. The anchorperson is the one who appears the most frequently in each news video. Hence, we apply the densest component algorithm to each video separately and automatically detect the anchorperson, since the faces of the anchorperson will correspond to the densest component of the graph formed by the faces in that video.

1.3 Organization of the Thesis

The rest of the paper is organized as the following:

In Chapter 2, we summarize related previous works on the similar problems. in Chapter 3, we present a detailed explanation of the overall graph approach: association of names and faces to limit the search space, definition of similarity measure to construct the similarity graph, finding the largest densest component of the graph, and using the result in recognizing new faces. The method is also applied on news videos for detecting anchorpeople automatically. In Chapter 4, description of the datasets and experimental results are given. In Chapter 5, the proposed method is compared with other possible approaches. We finally summarize the study in Chapter 6 and conclude with future research.

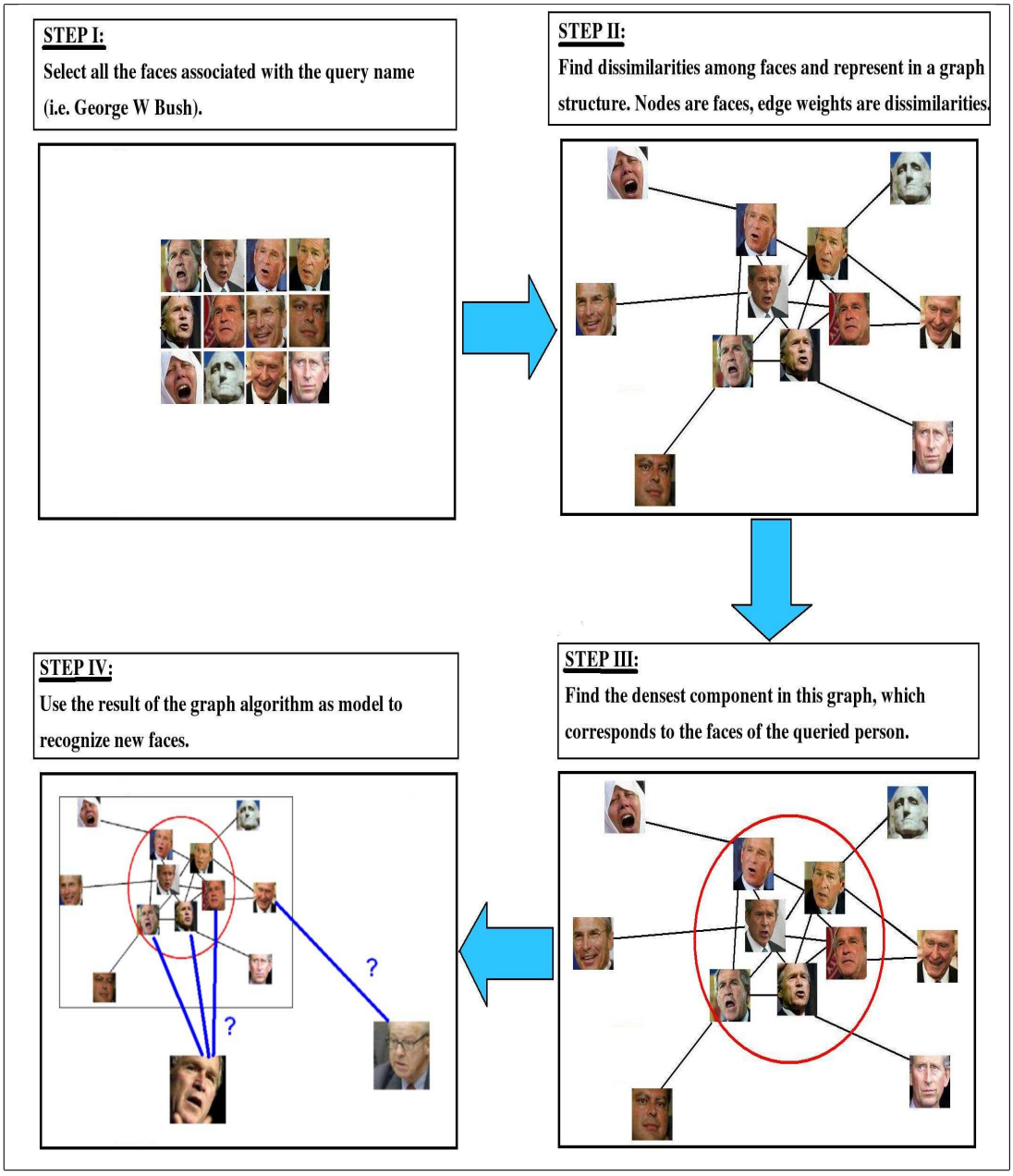


Figure 1.6: The four steps of overall approach: 1: Limit the search space by using name. 2: Construct a similarity graph among faces in this limited space. 3: Find the largest densest component of the graph corresponding to the faces of the query person. 4: Use the result as a model for recognizing new faces.

Chapter 2

Background

In this thesis, we propose a method for finding and recognizing faces in news by integrating names and faces with a graph based approach. In the following sections, we first discuss some of the previous work on the use of name and text information. Then, we define the problem of face recognition and importance of interest points in literature on solving the recognition problem. Later in the last section, we briefly discuss the use of graph theoretical methods in computer vision.

2.1 On Integration of Names and Faces

News video and photograph collections possess different types of resources, such as text, speech, and visuality. Recently, it has been shown that effective and efficient accessing and utilization of such multi-modal data can be simplified by integrating the different types of resources. In the most of previous work, names and faces are associated for better query results.

In [52], Yang et al. showed that the combination of text and face information allows better retrieval performances in news videos. In their work, they avoid the difficulties of face recognition problem by using the text information for selecting

some shots as the initial results and applying face recognition on those shots. They aim to reduce the number of faces to be recognized, hence improve the accuracy. The timing between names and appearances of people is modeled by propagating the similarity scores from the shots containing the query person's name to the neighboring shots in a window. The Eigenface algorithm is used for face recognition. In the paper, an anchor detector has also been built by combining three information resources: color histogram from image data, speaker ID from audio data, face info from face detection. They use Fisher's Linear Discrimant (FLD) to select the distinguishing features for each source of data. Then, they combine the selected features into a new feature vector to be used in classification. Upon a similar approach in [14], text and image features are used together to iteratively narrow the search for browsing and retrieving web documents. [12] also unifies the textual and visual statistics in a single indexing vector for retrieval of web documents.

Berg et al. [8] proposed a method for associating the faces in the news photographs with a set of names extracted from the captions. In that paper, they first perform kernel principal component analysis (kPCA) to reduce the dimensionality of the image data and linear discriminant analysis (LDA) to project the data into its discriminant coordinates. Each image is then represented with both a vector gained after the kPCA and LDA processes, and a set of associated names extracted from the caption. A modified version of k-means clustering is used to assign a label for each image. Clusters that are far away from the mean are removed from the data set, and discriminant coordinates are re-estimated. Finally, the clusters, which show high facial similarity are merged. The results given in this work, are then improved in [7] by analyzing language more carefully. In the latter work, they also learn a natural language classifier that can be used to determine who is pictured from text alone.

[18] integrates names and faces using speech transcripts, and improves the retrieval performance of person queries on TRECVID2004. It first searches over the speech transcript text and selects the key-frames that are aligned with the query name. Then, it applies Schneiderman-Kanade's face detection algorithm on each key-frame. However, many false positives are produced with such an

approach. Hence, skin color information is used to eliminate the false positives. The probability of a pixel being a skin pixel is modeled using a Gaussian probability distribution on HSV color space. Then, three features (color feature, PCA, ICA) are extracted from the faces to be use in grouping similar faces. G-means clustering is used in grouping, which starts from small number of clusters, K, and increases K iteratively if some clusters fail the Gaussianity test (Kolmogorov-Simirov test) [24]. After grouping, each cluster is represented by one representative face: the one that is the closest to the mean of its cluster. Hence, the proposed method increases the speed of the system by reducing the number of images provided to the user. Anchor filtering is also experimented by selecting the anchor representatives and removing them from the rest of the cluster.

In [2], a method is employed for automatically labeling of the characters in TV or film by using both visual and textual information. First, the subtitles and the script are aligned for tagging subtitles with speaker identity. Faces are detected and then tracked to compose face tracks. While obtaining the face tracks, the number of point tracks which pass through faces is counted, and if this number is large relative to the number of point tracks which are not in common to both faces, a match is declared. To represent the appearance of a face, nine facial features are located and two local feature descriptors are used: sift and simple pixel-wised descriptor formed by taking the vector of pixels in the elliptical region and normalizing to obtain local photometric invariance. Further to use the additional cues, a bounding box is predicted for each face, which is expected to contain the clothing of the corresponding character. Speaker are detected by a simple lip motion detection algorithm. Then, each unlabeled face track is represented by a set of face and clothing descriptors. Finally, a probabilistic model is used to classify these tracks based on the weighted probabilities of the face and cloth appearance features.

2.2 On Face Recognition

Face recognition is a long-standing and a difficult problem, which has received great attention due to its demand on different fields like commercial and law enforcement applications. The problem has aroused interest in various science domains such as pattern recognition, computer vision, image analysis, psychology, and neurosciences. Findings in any one of those science has a direct effect on the others. As stated in [54], there exists still many ambiguous questions involved in the process of human perception of faces that psychologists and neuroscientists work on. Some of these questions are:

- Is face recognition a unique process in the brain and different from the process of object recognition?
- Do we recognize faces as a whole or by individual features?
- Which features are more important for face recognition?
- Is human face perception invariant to changes in view-point, lighting and/or expressions?

In the context of computer vision, a broad statement of the face recognition problem is declared as: Given a still or a video image, identify or verify one or more persons in the image by using a stored database of faces([54, 49]). There are three main steps involved in the configuration of a generic face recognition system (see Figure 2.1): detection of the face region in the image, feature extraction from the face region, recognition.

In [54] and [47], the face recognition methods have been put into one of the three categories: holistic methods, local methods or hybrid methods. We give a brief summary of the those three approaches in the next three subsections.

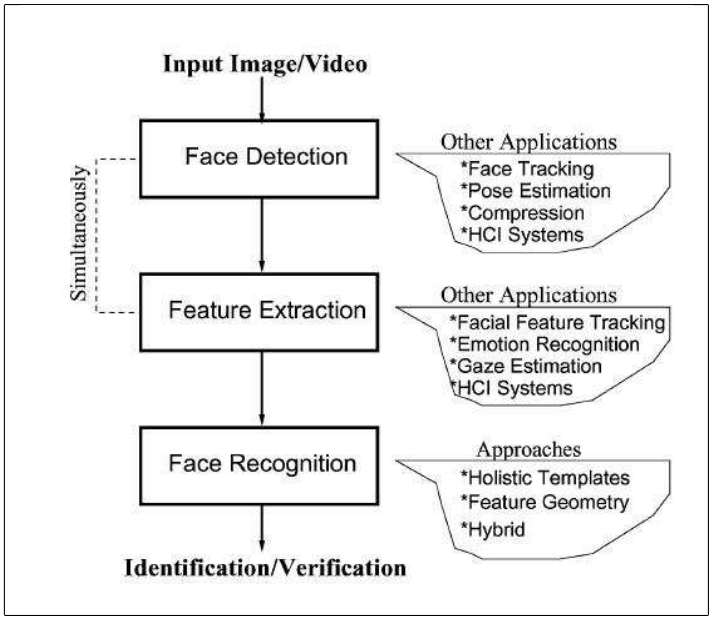


Figure 2.1: The main steps in the face recognition process (taken from [54]).

2.2.1 Holistic Methods

Holistic methods use the whole face as the raw input to the system. In those methods, each face image is represented with a vector composed by concatenating the gray-level pixel values of the image. Among the holistic methods, the most popular techniques used are Principal Component Analysis (PCA), Fisher Discriminant Analysis (FDA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA).

The main idea in PCA is to project the training data into a sub-dimensional feature space, in which basis vectors correspond to the maximum variance direction in the original image space. Each PCA basis vector was referred as an "eigenface" in [35, 36] that can be displayed as a sort of ghostly face. In the classification stage, the input face image is first projected into the subspace spanned by the eigenfaces; where each face is represented by a linear combination of the eigenfaces. Then, the new face is classified according its position in the face space to the positions of the known individuals. Later in literature, extensions of the Principal Component Analysis have been proposed as in [53] by using two-dimensional PCA and in [50] by selecting discriminant eigenfaces for face

recognition.

In [4], Bartlett et al. claimed that ICA representations of faces are superior to PCA; hence ICA can perform better across sessions and changes in expression. In ICA, data is projected onto some basis that are statistically independent. It also attempts to minimizes second-order and higher-order dependencies in the input data.

Holistic methods can be advantageous in the context of covering the global appearance of the faces. One other edge of the holistic methods is that they maintain the diffuse texture and shape information; hence can differentiate the faces. However, representing each face image by a feature vector makes the holistic methods sensitive to variations in appearance caused by occlusion, changes in illumination and/or expressions. To overcome this problem, local feature representations have been developed in literature that are less sensitive to changes in appearance.

2.2.2 Local Methods

The importance of hair, eyes and mouth in face human perception has been highlighted in psychology and neurosciences ([27, 10]). It has been claimed in [10] that nose plays an important role in face recognition. On these grounds, local methods have been presented that represent each face image by a set of low dimensional feature vector, which usually correspond to the facial features like eyes, mouth, and nose.

Mainly, there are two approaches used in local methods: local feature-based methods and local appearance-based methods. The geometric relations of the features are considered in local feature-based methods. In some primitive studies of this types, only the geometric measures –such as distance between the eyes or the size of the eyes– have been considered [30, 29]. Then, in [37] Manjunath et al. proposed a method that preserves both the local information and the global topology of the face. The method stores both the local information and the

feature information of the detected facial features; and constructs a topological graph using these features. Then, the classification stage, the problem is solved by a graph matching scheme.

Even if the method in [37] is advantageous since it considers both the similarity of the local features and the global topology, it is sensitive variations that causes any change in this topology. Consequently, Lades et al. [11] proposed Elastic Bunch Graph Matching Scheme (EBGM) that is based on a deformable topology graph. Even though the method resolves the problem caused by the changes in appearance, it cannot remove the problem caused by the occlusion of any of facial features.

Motivated by the inventions in psychology that a set of simple lines can characterize the structure of an object, hence is sufficient to recognize its shape; a face is represented by the Face-ARG in [42]. Face-ARG consists of a set of nodes that corresponds to line features, and binary relations between them. Using a Face-ARG, all the geometric quantities and structural information of the face can be encoded in an Attributed Relational Graph (ARG) structure. Then, in the classification stage, partial ARG matching is applied on the constructed Face-ARG's of the test and reference faces. The advantage of the method over recent methods is that it can still work in presence of occlusion and expression changes.

Most of the local feature-based methods are sensitive to accurate feature point localization; which is still an unresolved problem. Hence, in local appearance-based methods, facial feature regions are detected first and features (such as Gabor wavelet [37] and Haar wavelet [33]) are extracted from these regions in the image. Different classifiers are applied on these features finally in the classification step.

2.2.3 Hybrid Methods

It is still a question if we recognize faces as a whole or by individual features. Holistic methods cover the global face appearance and can maintain the diffuse

texture and shape information. However, they are very sensitive to the changes in appearance that may be caused by occlusion, changes in illumination and/or expressions. Local methods is less sensitive to those changes, however accurate localization of facial feature points still remains that can affect the performance of recognition. Especially for images of small size, facial feature detection is even more problematic; hence the local methods cannot be applied in that case. How to use the local features to without disregarding the global structure is another open problem. To overcome those problems and benefit from different advantages of both approaches, hybrid methods can be used. Hybrid methods aim to use both holistic and local approaches. However, how to use which features and which classifiers in hybrid methods is still indeterminate and not much investigated in literature yet.

2.3 On the Use of Interest Points

Recently, it has been shown that local image features provide a good representation of the image for recognition and retrieval [5, 40], while global features tend to be sensitive to image variations in viewpoint, pose and illumination. Among the local features, Scale Invariant Feature Transform (SIFT) technique has been shown to be successful in recognizing objects [34, 38, 25] and faces [26, 9]. The technique, which has first been presented in [34], provides distinctive invariant features that can be used in reliable matching between different images of an object or scene. The most powerful aspect of these features is that they are invariant to image scale and rotation, and partially invariant to change in illumination and 3D camera viewpoint.

In [9], application of the SIFT approach in the scope of face authentication has been invested. There different matching schemes are proposed in the paper: 1. Minimum pair distance, 2. Matching eyes and mouth, 3. Matching on a regular grid. The first scheme proposes to compute the distances between all pairs of key-points in two images and assign the minimum distance as the matching score. Using the that face and mouth regions provide most of the information for face

recognition, only the features in these regions are used in the second scheme. Finally, feature locations are considered in the third scheme by dividing the face image into grids and matching the features of corresponding grids.

Sivic et al. has proposed a person retrieval system in [26] that represents each face image as a collection of overlapping local SIFT descriptors placed at the five facial feature locations (eyes, mouth, nose, and middle point between the eyes). They first use tracking to associate faces into face-tracks within a shot to obtain multiple exemplars of the same person. Then, they represent each face-track with a histogram of precomputed face-feature exemplars. This histogram is used for matching the face-tracks; hence retrieving sets of faces across shots.

2.4 On the Use of Graph Theoretical Methods in Computer Vision

Graph theoretical approaches have recently been used in computer vision problems due to their representational power and flexibility. They allow vision problems to be cast in a strong theoretical area and access to the full depot of graph algorithms developed in computer science and operations research. The most common graph theoretical problems used in computer vision include maximum flow, minimum spanning tree, maximum clique, shortest path, maximal common subtree/ subgraph, graph partitioning, graph indexing, graph matching, etc. [17].

Graph partitioning algorithms that have been used in [21, 28, 45], target the two typical applications of computer vision: image segmentation and perceptual grouping. They address the problem of making cuts in a weighted graph according to an appropriate minimum weight criterion. In these works, data elements (i.e. image pixel points) correspond to the vertices in the graph, and similarity between any two vertices correspond to the edge weight between those vertices. In [6], the problem of content based image retrieval has been solved by a graph matching scheme. The main idea used was to represent an image query as an attributed relation graph, and select a small number of model image graphs that are similar

to the query image graph.

[3] has proposed a graph theoretic clustering method for image grouping and retrieval. The motivation of the work was that an efficient retrieval algorithm should be able to retrieve images that are not only close (similar) to the query image but also close to each other. However, most of the existing feature extraction algorithms cannot always map visually similar images to nearby locations in the feature space. Hence in the retrieval step, it is often to retrieve irrelevant images (or not to retrieve relevant images) simply because they are close to the query image (or a bit far away from the query image). In this context, they retrieve best N matches for a query image, and best N matches of each of the retrieved images. A graph is constructed with all those retrieved images, in which nodes corresponds to the images and edge weights correspond to the similarities. Then, the retrieval problem is transformed into and solved by the problem of finding the set of nodes in the graph, that are not as dense as major cliques but are compact enough within user specified thresholds.

Chapter 3

Graph Based Person Finding Approach

3.1 Overview

It is likely that in the news, a person will appear when his/her name is mentioned. Following up this cue, we start search for a person by first looking for the name of the query person and limit our search space to the images that are associated with that name. Although there might be the faces of other people or non-face images in this limited search space, mostly query person will be the one that appears more than any other individual. Visually, faces of a particular person tend to be more similar to each other than to faces of other people. Based on these assumptions, we transform the problem to a graph problem, in which the nodes correspond to faces and the edges correspond to the similarities between faces, and we seek to find the largest densest component in this graph. Hence, if we can define a similarity measure among the faces in the limited search space and represent the similarities in a graph structure, then the problem of finding the most similar faces corresponding to the instances of query name's face can be tackled by finding the densest component in the graph.

In the following sections, we first explain the steps of the graph based person

finding approach: associating names and faces to limit the search space, defining a similarity measure between faces to construct the similarity graph, and the greedy graph algorithm to find the largest densest component of the graph. Following those sections, we explain the use of person finding approach for automatic anchorperson detection in news videos. Then, in the last section we present two methods for recognizing new faces by using the output of the graph based person finding approach.

3.2 Integrating Names and Faces

The first step of our algorithm involves associating name and face information. In this step, we use the name information to limit our search space to the images around which the name of the query person appears. To this end, we look for the name of the query person in the caption or in the speech transcript; and choose the images that are associated with the query name.

On the web, the news photographs appear with the captions. Using the assumption that a person is likely to appear in a photograph when his/her name is mentioned in the caption, we reduce the face set for a queried person by only choosing the photographs that include the name of that person in the associated caption. However, a person's name can appear in different forms. For example, the names *George W. Bush, President Bush, U.S. President*, and *President George Bush*, all correspond to the same person. We merged the set of different names used for the same person to find all faces associated with all different names of the same person. A detailed list of different forms of names corresponding to each query person used in our experiments is given in the appendix.

For the news videos, the probability of a person appearing on the screen is high when his/her name is mentioned in the speech transcript text. Thus, looking for the shots in which the name of the query name is mentioned is a good place to start search over people. However, it is problematic since there can be a time shift between the appearance of the person visually and the appearance of his/her

name.

Recently, it is shown that the frequency of a person's visual appearance with respect to the occurrence of his/her name can be assumed to have a Gaussian distribution [52]. We use the same idea and search for the range where the face is likely to appear relative to the name. As we experimented on "Clinton" query, we see that taking the ten preceding and succeeding shots together with the shot, where the name is mentioned is a good approximation to find most of the relevant faces (see Figure 4.9). However, the number of faces in this range (which we refer to as [-10,10]) can still be large compared to the instances of the query name. As we will explain in the experiments, it is seen that taking only one preceding and two following shots (which we refer to as [-1,2]) is also a good choice.

Another problem in news videos is that usually the faces of the anchorperson appear around a name. For solution to this problem, we use an anchorperson detection method based on our graph based approach as we will be explained later.

As mentioned previously, photographs/shots associated with the name may not include any people. Thus to eliminate such cases, we apply a face detection algorithm on the retrieved retrieved images. However, there can be more than one face in the image and more than one name in the corresponding text/speech. Therefore, it is not known which face goes with which name.

Integrating names and faces produces better retrieval performances compared to solely text-based methods. However, the resulting set may still contain many false faces due to the following reasons: the query person may not appear visually even if his/her name is mentioned, there may be other people in the same story that also appear visually together with the query person, there may be non-face images returned by the face detection algorithm used. However, the faces of the query person are likely to be the most frequently appearing ones than any other person in the same space. Even if the expressions or poses vary, different appearances of the face of the same person tend to be more similar to each other than to the faces of others. In the following sections, we explain our strategy to increase retrieval performance by finding the correct faces by using visual

similarities.

3.3 Constructing Similarity Graph of Faces

We represent the faces with the interest points extracted from the images using the SIFT operator [34]. Lowe's SIFT operator [34], have recently been shown to be successful in recognizing objects [38, 25] and faces, [26]. The SIFT technique consists of four main steps: 1. Scale-space extrema detection, 2. Keypoint localization, 3. Orientation Assignment, 4. Keypoint descriptor.

In the first step, potential interest points are extracted by looking for all scales and image locations. A scale space of the image is constructed first by convolving the image with variable-scale Gaussian function. Let the input image be $I(x, y)$ then, the Gaussian-blurred image $L(x, y, \sigma)$ can be represented by:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y),$$

where,

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

Difference-ofGaussian function (DoG) is used to detect stable keypoint locations in this scale space. DoG of two nearby scales separated by constant k is given by:

$$\begin{aligned} DoG(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma). \end{aligned}$$

Local maxima and/or minima of DOF gives the candidate keypoint location, and is computed by comparing each sample point with both the eight neighbor points on the same scale and the nine neighbor points on two images in the neighbor scale.

In the second step, each candidate keypoint location is fit to a nearby data to determine its location, scale and ratio of principal curvatures. First, a threshold on minimum contrast is applied to remove the unstable extrema with low contrast. Then, a second threshold is applied on the ratio of principle curvatures to eliminate the strong responses of difference-of- Gaussian function along edges which are poorly determined.

An orientation is assigned to the keypoint locations in the third step, based on local image gradient directions. An orientation histogram is formed from gradient orientations of the sample points around a keypoint by precomputing their pixel differences in a scale invariant manner. Peaks in the histogram denominate dominant directions of local gradients; thus the 80 per cent of the highest peak in the histogram is used to assign the orientation of a keypoint.

In the last step, a descriptor for each keypoint is computed from image gradients. Gradients of points within an array around the keypoint are weighted by a Gaussian window and its content is summarized into a descriptor array, using orientation histograms. These features are then normalized to unit length and a threshold is applied to this unit vector values to reduce the effects of illumination change.

We first use a minimum distance metric to find all matching points and then remove the false matches by adding some constraints. For each pair of faces, the interest points on the first face are compared with the interest points on the second face and the points having the least Euclidean distance are assumed to be the correct matches. However, among these there can be many false matches as well (see Figure 3.1).

In order to eliminate the false matches, we apply two constraints: the geometrical constraint and the unique match constraints. Geometrical constraint expects the matching points to appear around similar positions on the face when the normalized positions are considered. The matches whose interest points do not fall in close positions on the face are eliminated. Unique match constraint ensures that each point matches to only a single point by eliminating multiple matches to one point and also by removing one-way matches. In the next two

subsections, we give the details of how those constraints are applied.

3.3.1 Geometrical Constraint

We expect that matching points will be found around similar positions on the face. For example, the left eye usually resides around the middle-left of a face, even in different poses. This assumption presumes that the matching pair of points will be in close proximity when the normalized coordinates (the relative position of the points on the faces) are considered.

To eliminate false matches which are distant from each other, we apply a geometrical constraint. For this purpose, we randomly selected 5 images of 10 people. Then, we manually assigned true and false matches for each comparison and used them as training samples to be run on a quadratic Bayes normal classifier ([43, 51]) to classify a matched point as true or false according to its geometrical distance. The geometrical distance corresponding to the i^{th} assignment refers to $\sqrt{X^2 + Y^2}$ where

$$X = \frac{locX\,(i)}{sizeX\,(image1)} - \frac{locX\,(match\,(i))}{sizeX\,(image2)},$$

$$Y = \frac{locY\,(i)}{sizeY\,(image1)} - \frac{locY\,(match\,(i))}{sizeY\,(image2)},$$

and $locX$ and $locY$ hold X and Y coordinates of the feature points in the images, $sizeX$ and $sizeY$ hold X and Y sizes of the images and $match(i)$ corresponds to the matched keypoint in the second image of the i^{th} feature point in the first image.

In Figure 3.1, matches before and after the application of this geometrical constraint are shown for an example face pair. Most of the false matches are eliminated when the points that are far away from each other are removed.

The relative angle between the points could also be used as a geometric constraint. However, since the closer points could have very large angle differences,

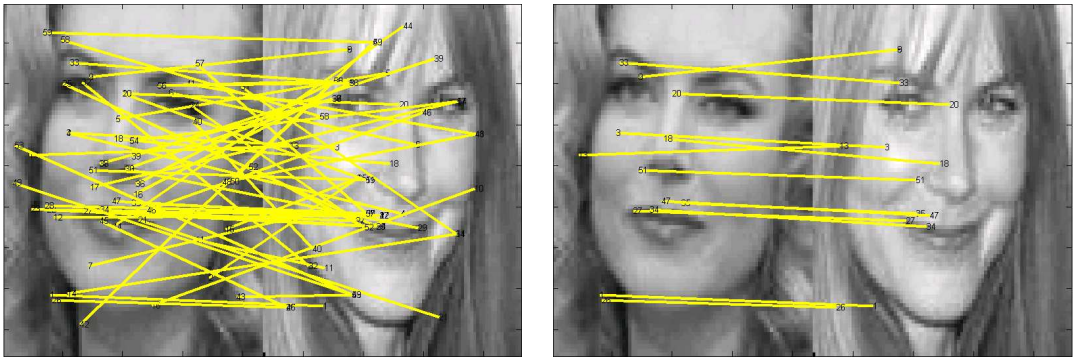


Figure 3.1: The first image on the left shows all the feature points and their matches based on the minimum distance. The second image on the right shows the matches that are assigned as true after the application of geometrical constraints.

it is not reliable.

3.3.2 Unique Match Constraints

After eliminating the points that do not satisfy the geometrical constraints, there can still be some false matches. Usually, the false matches are due to *multiple assignments* which exist when more than one point (e.g, A_1 and A_2) are assigned to a single point (e.g, B_1) in the other image, or to *one way assignments* which exist when a point A_1 is assigned to a point B_1 on the other image while the point B_1 is assigned to another point A_2 or not assigned to any point (Figure 3.2). These false matches can be eliminated with the application of another constraint, namely the unique match constraint, which guarantees that each assignment from an image A to another image B will have a corresponding assignment from image B to image A.

The false matches due to multiple assignments are eliminated by choosing the match with the minimum distance. The false matches due to one way assignments are eliminated by removing the links which do not have any corresponding assignment from the other side. An example showing the matches before and after applying the unique match constraints are given in Figure 3.3 and in Figure 3.4.

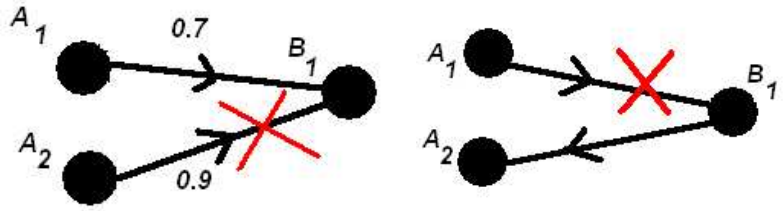


Figure 3.2: For a pair of faces A and B, let A_1 and A_2 be two points on A; and B_1 is a point on B with the arrows showing the matches and their direction. On the left is a *multiple assignment* where both points A_1 and A_2 on A match B_1 on B. In such a case, the match between A_2 and B_2 is eliminated. On the right is a *one way match* where B_1 is a match for A_1, whereas B_1 matches another point A_2 on A. The match of A_1 to B_1 is eliminated. The match of B_1 to A_2 remains the same if B_1 is also a match for A_2; otherwise it is eliminated.

3.3.3 Similarity Graph Construction

After applying the constraints and assuming that the remaining matches are true matches, we define the distance between the two faces A and B as the average value of all matches.

$$dist(A, B) = \frac{\sum_{i=1}^{N} D(i)}{N},$$

where N is the number of true matches and $D(i)$ is the Euclidean distance between the SIFT descriptors of the two points for the i^{th} match.

A similarity graph for all faces in the search space is then constructed using these distances. We can represent the graph as a matrix as in Figure 3.5. The matrix is symmetric and the values on the diagonal are all zero. For a more clear visual representation, the distances for the faces corresponding to the person we are seeking are shown together. Clearly, these faces are more similar to each other than to the others. Our goal is to find this subset which will correspond to the densest component in the graph structure.

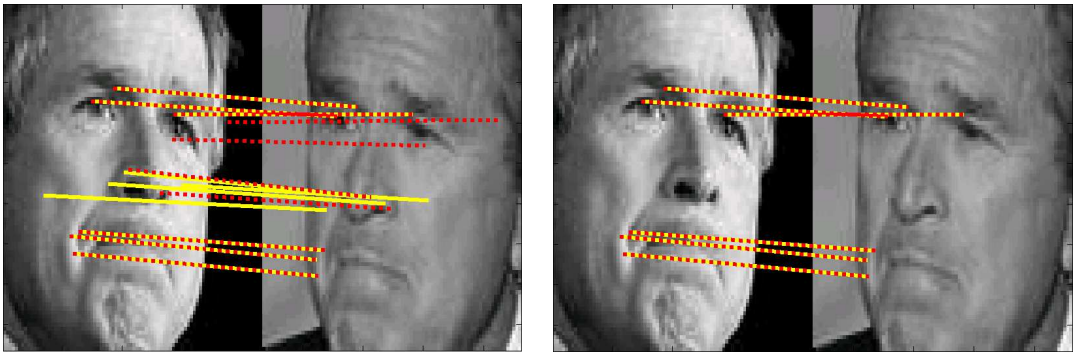


Figure 3.3: An example for unique match constraint. Matches from the left to the right image are shown by red, dashed lines, whereas matches from right to left are shown by yellow lines. The left image shows the matches assigned after applying geometrical constraints, but before applying the unique match constraints. The right image shows the remaining matches after applying the unique match constraints.

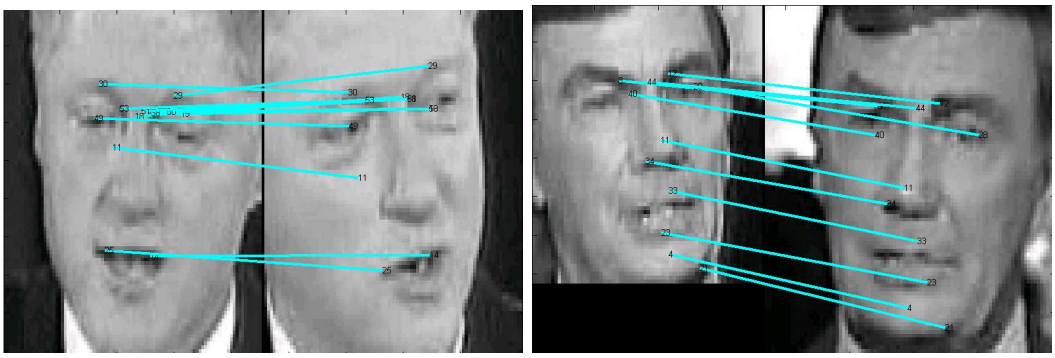


Figure 3.4: Sample matching points from news videos dataset after applying all the constraints. Note that, even for faces with different size, pose or expressions the method successfully finds the corresponding points.

3.4 Greedy Graph Algorithm for Finding the Densest Component

In the constructed similarity graph, faces represent the nodes and the distances between the faces represent the edge weights. We assume that, in this graph the nodes of a particular person will be close to each other (highly connected) and distant from the other nodes (weakly connected). Hence, the problem can be transformed in to finding the densest subgraph (component) in the entire graph. To find the densest component we adapt the method proposed by Charikar [13]

(a) (b)

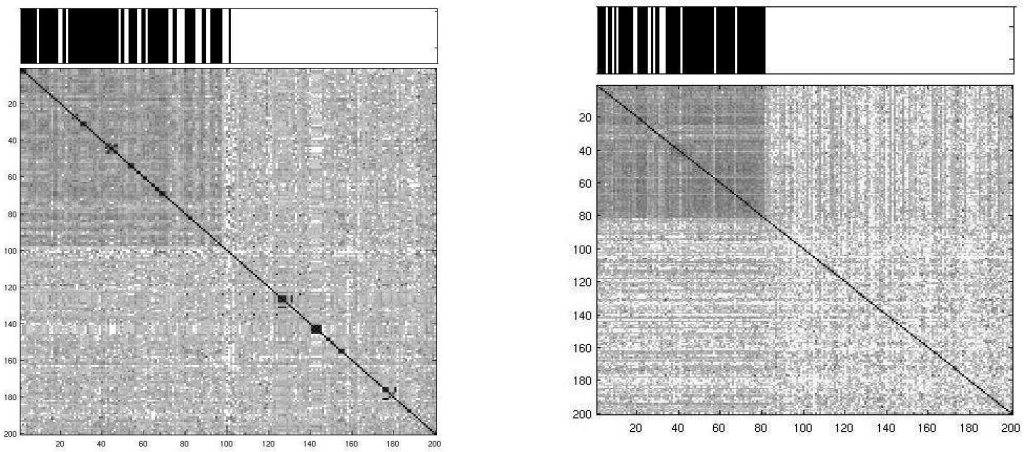


Figure 3.5: Samples for constructed similarity matrices. For visualization, the true images are put on the top left of each matrix. Dark colors correspond to larger similarity values. Bars on the top of the matrices indicate the items that are assigned to be in the largest densest component of the corresponding graph. (a) Similarity matrix for 200 images in the search space for the name *Hans Blix*. In this search space, 97 of the images are true *Hans Blix* images, and the remaining 103 are not. (b) Similarity matrix for 200 images in the search space for the name *Sam Donaldson*. 81 of the images are true *Sam Donaldson* images, and the remaining 119 are not.

where the density of subset S of a graph G is defined as

$$f(\ S) = \frac{\mid E(\ S)\mid}{\mid S\mid},$$

where $E(\ S) = \{i, j \in E : i \in S, j \in S\}$ and E is the set of all edges in G. In other words, $E(S)$ is the set of edges induced by subset S. The subset S that maximizes $f(S)$ is defined as the densest component.

Our goal is to find the subgraph S with the largest average degree that is the subgraph with the maximum density. Initially, the algorithm presented in [13] starts with the entire graph G and sets $S = G$. Then, in each step, the vertex with the minimum degree is removed from S. The algorithm also computes the value of $f(S)$ for each step and continues until the set S is empty. Finally, the set S, that has maximum $f(S)$ value, is returned as the densest component of

the graph.

In order to apply the above algorithm to the constructed similarity graph, we need to convert it into a binary form, since the algorithm described above only works well for binary graphs. Thus, before applying it, we convert our original dissimilarity values into a binary form, in which 0 indicates no edge and 1 indicates an edge between two nodes. This conversion is carried out by applying a threshold on the distance between the nodes. This threshold also connotes what we define as near-by and/or remote. An example of such a conversion is given in Figure 3.6. In the example, assume that 0.65 is defined as our proximity threshold. In other words, if the distance between two nodes is less than or equal to 0.65 then these two nodes are near-by; therefore we put an edge between them. Otherwise, no edge is maintained between these nodes, since they are far away from each other.

3.5 Anchorperson Detection and Removal for News Videos

When we look at the shots where the query name is mentioned in the speech transcript, it is likely that the anchorperson/reporter might be introducing or wrapping up a story, with the preceding or succeeding shots being relevant, but not the current one. Therefore, when the shots including the query name are selected, the faces of the anchorperson will appear frequently making our assumption that the most frequent face will correspond to the query name wrong. Hence, it is highly probable that the anchorperson will be returned as the densest component by the person finding algorithm (see Figure 3.7). The solution is to detect and remove the anchorperson before applying the algorithm

In [15], a supervised method for anchorperson detection is proposed. They integrate color and face information together with speaker-id extracted from the audio. However, this method has some disadvantages. First of all, it highly depends on the speaker-id, and requires the analysis of audio data. The color information is useful to capture the characteristics of studio settings where the

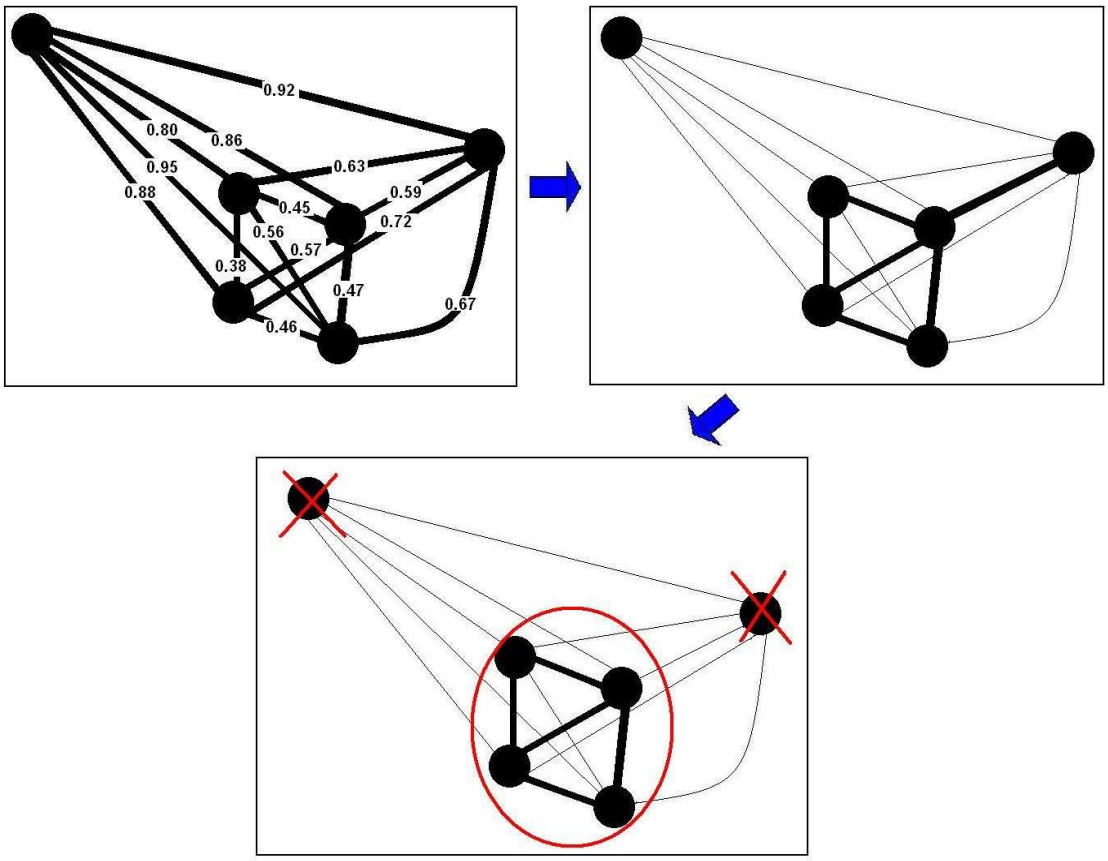


Figure 3.6: Example of converting a weighted graph to a binary graph. Nodes and their distances are given in the first image. The resulting graph after applying 0.65 as the proximity threshold is given in the second image. Bold edges are the edges that remain after conversion. The final densest component of this graph is circled in the last image.

anchorperson is likely to appear. But, when the anchorperson reports from another environment this assumption fails. Finally, the method depends on the fact that the faces of anchorpeople appear in large sizes and around some specific positions, but again there may be cases where this is not the case.

In this study, we use the graph based method to find the anchorpeople in an unsupervised way. The idea is based on the fact that, the anchorpeople are usually the most frequently appearing people in broadcast news videos. For different days there may be different anchorpeople reporting, but generally there is a single anchorperson for each day. Hence, we apply the densest component based method to each news video separately, to find the people appearing most

frequently, which correspond to the anchorpeople.

3.6 Dynamic Face Recognition

The overall scheme explained in the previous sections returns a set of images classified as the query person (densest component) and the rest as others (outliers). Also, the graph algorithm works on the whole set of images in the search space of the query person. Thus, when a new face is encountered, the algorithm needs to be re-run on the whole set to learn the label of the new face–query person or outlier. However, since the scheme returns the classified images, this result can be used as a model to recognize new faces dynamically and check if it belong to the faces of the queried person. Moreover, this task can be achieved without any supervision, since the scheme provides us the training data labeled automatically.

In the next two subsections, we explain how the output of the person finding approach can be used in recognizing new faces. We model the returned solution in two ways to learn the thresholds for: average degree and average distance.

3.6.1 Degree Modeling

As explained in 3.4,the greedy densest component algorithm works iteratively by removing one node from the graph until there is one last node left in the graph. Average density of each subgraph is computed in each iteration and finally the subgraph with the largest average density is assigned as the densest component. Among these iterations, there is one *lastnode* removed from the current subgraph that results in the densest component in the next iteration. This last node can be thought of as the breaking point, and indicate an evidence for the maximum number of total connections (edges) from an outlier node to all the nodes in the densest component. This total number–degree of the nearest outlier to all the faces recognized as the query person–can used as a threshold in further recognition. When a new face is encountered, its degree to all the faces in

the densest component is computed first. Then, the face is labeled as the query person if its degree is greater than the found degree threshold.

3.6.2 Distance Modeling

In this method, average distance of true-true and false-true matches are used. For each node in the graph, its average distance to all the nodes in the densest component– hence the faces of the query person–are computed. If a node was in the densest component, then its average distance is labeled as a true-true match distance, else a false-true match distance. These distances are then trained with the quadratic Bayes normal classifier ([43, 51]) to learn the average distance threshold and classify new test images based on its average distance to true images in the training set.

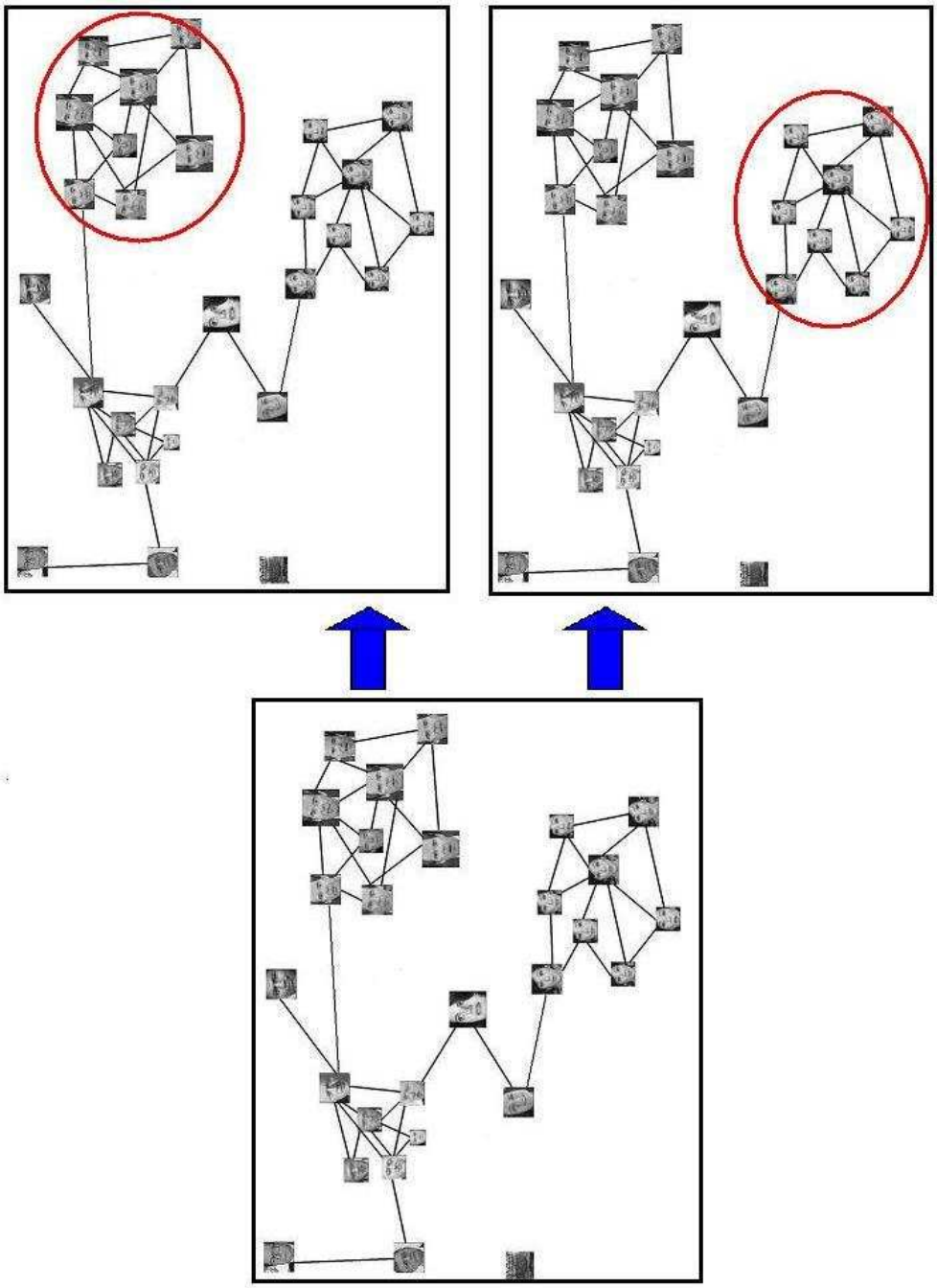


Figure 3.7: The first figure on the left corresponds to a representative limited search space for the query name *Bill Clinton*. It is highly probable that one of the anchorpeople will form the largest densest component in this limited search space. Hence, either one of them will be returned as the query person like in the figures on the right.

Chapter 4

Experiments

4.1 Datasets

The method proposed in this thesis is tested on two different datasets: news photographs on the web and broadcast news videos on television. In the next two subsections, we briefly describe both datasets used in our experiments.

4.1.1 News Photographs

The dataset constructed by Berg *et al.* originally consists of about half a million captioned news images collected from Yahoo! News on the Web. After applying a face detection algorithm and processing the resulting faces, they were left with a total of 30,281 detected faces [7]. Each image in this set is associated with a set of names. A total of 13,292 different names are used for association. However more than half (9,609) of them are used only once or twice. Also, as we mentioned previously, a particular person may be called by different names. For example, the names used for *George W Bush* and their frequency are: *George W (1485); W. Bush (1462); George W. Bush (1454); President George W (1443); President Bush (905); U.S. President (722); President George Bush (44); President Bushs (2); President George W Bush (2); George W Bush (2).* We merge the set of

different names used for the same person and then take the intersection to find faces associated with different names of the same person. A detailed list of different forms of names corresponding to each query person used in our experiments is given in the appendix.

Generally, the number of faces in the resulting set is less than the number of all names since a caption may include more than one instance of the referred name. For example, for *Bush* the number of faces is 2,849 while the total number of all referred names is 7,528. In the experiments, the top 23 people appearing with the highest frequencies (more than 200 times) are used. Figure 4.1 shows the total number of faces associated with the given name and the number of correct faces for the 23 people used in the experiments.

4.1.2 News Videos

The second dataset used in the experiments is the broadcast news videos provided by NIST for TRECVID video retrieval evaluation competition 2004 [1]. It consists of 229 movies (30 minutes each) from ABC and CNN news. The shot boundaries and the key-frames are provided by NIST. Speech transcripts extracted by LIMSI [20] are used to obtain the associated text for each shot.

For the experiments, we choose 5 people, namely Bill Clinton, Benjamin Netanyahu, Sam Donaldson, Saddam Hussein and Boris Yeltsin. In the speech transcript text, their names appear 991, 51, 100, 149 and 78 times respectively.

The face detection algorithm provided by Mikolajcyzk [39] is used to extract faces from key-frames. Due to high noise levels and low image resolution quality, the face detector produces many false alarms. On randomly selected ten videos, in 2942 images, 1395 regions are detected as faces but only 790 of them are real faces and 580 faces are missed. In total, 31,724 faces are detected over the whole dataset.

4.2 Evaluation Criteria

For evaluation, we give the experimental results based on recall and precision values. These values are computed as follows: Let N be the total number of faces returned by the algorithm as the faces of the query person. Among those N, let n be the total number of faces that really belong to that person. Then, the precision value of this result is:

$$precison = \frac{n}{N}.$$

If there is a total of m faces in the whole dataset that belong to the query person, then the recall value of the result is:

$$recall = \frac{n}{m}.$$

We should also denote that as the baseline, we use the images returned by the face detector. Hence, m (ground truth of the query person) is computed among those detected faces.

After finding the recall and precision values for each query person individually, we finally compute the weighted results for average recall and precision values. What we mean with weighted is that recall and/or precision value of each person is weighted by the number of images that appear in his/her limited search space. Let $recall(i)$ be the recall value for the i^{th} person, and $S(i)$ be the total number of images in its limited search space. Then, weighted average recall is:

$$weightedAvgRecall = \frac{\sum S(i) * recall(i)}{\sum S(i)}.$$

All the average results given in the rest of the paper refers to the weighted averages.

In the following two subsections, we give the results of these experiments on both sets separately.

4.3 Experimental Results on News Photographs

4.3.1 Matching Points

As the first step, the points having the minimum distance according to their SIFT descriptors are defined as the matching points. These points are further eliminated using the two constraints. After this elimination process, 73% of all possible true matches are kept and we lose only 27% of true matches. Among these assignments, we achieved a correct matching rate of 72%.

4.3.2 Graph Approach

The success of our algorithm varies with the threshold that is chosen while converting the weighted dissimilarity graph to a binary one. To show the effect, average recall and precision values are plotted as in Figure 4.3 for varying thresholds between 0.55 and 0.65. Based on these values, the threshold 0.575 is chosen to represent the recall and precision values for each person.

Recall and precision values for each person using the threshold value of 0.575 are given in Figure 4.2. Average precision value is obtained as 48% for the baseline method which assumes that all the faces appearing around the name is correct. With the proposed, method we achieved 68% recall and 71% precision values on the average. The method can achieve up to 84% recall- as for *Gray Davis*- and 100% precision - as for *John Ashcroft, Hugo Chavez, Jiang Zemin and Abdullah Gul*. We had initially assumed that, after associating names, true faces of the queried person appear more than any other person in the search space. However, when this is not the case, the algorithm gives bad retrieval results. For example, there is a total of 913 images associated with name *Saddam Hussein*, but only 74 of them are true *Saddam Hussein* images while 179 of them are *George Bush* images.

To show that our system works also on individuals appearing in a small number of captions, we performed experiments on 10 people appearing less than 35 times and obtained average recall and precision values 85% and 66%. As another experiment, we changed the number of instances of a face by removing some of the correct faces or by adding some incorrect faces. For 4 people having around 200 instances and similar number of true and false images (i) we removed 50 of true images of from each of their search space, (ii) we added 100 false images. Originally, average recall and precision values were 63% and 95%. We obtained 59% recall an 89% precision after (i), and 58% recall and 70% precision after (ii). Although the precision is somewhat affected, results are still acceptable.

Some sample images retrieved and not retrieved for three people from the test set are shown in Figure 4.4.

4.3.3 Online Recognition

The recognition methods explained in 3.6 are tested on the news photographs dataset. In these experiments, K percentage of the images in the search space of a query person is selected as the held-out set and the graph algorithm is applied on the remaining images to learn the model. For each K, the algorithm is run 10 times with different set of random images for held-out set and model learning set. Average results of the degree modeling method are given in Figure 4.5 and in Table 4.1. And the average results of the distance modeling method are given in Figure 4.6 and in Table 4.2. For K = 10, the recall and precision values for each 23 person is also given in Figures 4.7 and 4.8 respectively for distance modeling technique.

Table 4.1: Recognition rates of degree modeling for different K values. (K per cent of the images are used for held-out.

K	10	20	30	40	50	60	70	80	90
model									
recall	0.68	0.68	0.67	0.66	0.66	0.65	0.63	0.61	0.58
precision	0.71	0.71	0.71	0.70	0.70	0.70	0.69	0.68	0.64
held-out set									
recall	0.69	0.70	0.70	0.70	0.69	0.69	0.69	0.70	0.76
precision	0.71	0.71	0.70	0.70	0.70	0.69	0.68	0.66	0.62

Table 4.2: Recognition rates of distance modeling for different K values. (K per cent of the images are used for held-out.

K	10	20	30	40	50	60	70	80	90
model									
recall	0.69	0.68	0.67	0.67	0.66	0.65	0.63	0.61	0.58
precision	0.71	0.71	0.70	0.70	0.70	0.70	0.69	0.68	0.65
held-out set									
recall	0.72	0.70	0.70	0.68	0.67	0.67	0.64	0.62	0.57
precision	0.72	0.72	0.72	0.72	0.72	0.72	0.71	0.71	0.69

4.4 Experimental Results on News Videos

4.4.1 Integrating Faces and Names

For better understanding of the distributions, we plot the frequency of faces relative to the position of the names for the five people that we have chosen for our experiments in Figure 4.10. It is seen that taking only one preceding and two following shots (which we refer to as [-1,2]) is also a good choice. Table 4.3 shows that, most of the correct faces fall into this selected range by removing many false alarms.

Table 4.3: Number of faces corresponding to the query name over total number of faces in the range [-10,10] and [-1,2].

Range	Clinton	Netanyahu	Donaldson	Saddam	Yeltsin
[-10,10]	213/6905	9/383	137/1197	18/1004	21/488
$[-1, 2]$	160/2457	6/114	102/330	14/332	19/157

4.4.2 Anchorperson Detection

We applied the densest component based method to each news video separately, to find the people appearing most frequently, which correspond to the anchorpeople. We run the algorithm on 229 videos in our test set, and obtained average recall and precision values as 0.90 and 0.85 respectively. Images that are detected as anchorperson in ten different videos are given in Figure 4.11.

4.4.3 Graph Approach

In order to determine a reasonable threshold used in converting the weighted similarity graph to a binary one for the news videos, we randomly selected 10 videos and recorded recall-precision values of different thresholds for anchorperson detection. These values are plotted in Figure 4.12. Further in our experiments, we select the point marked with a cross in the recall-precision curve, which corresponds to threshold of 0.6. The same threshold is used both for anchorperson detection and for person queries.

After selecting the range where the faces may appear we apply the densest component algorithm to find the faces corresponding to the query name. We have recorded the number of true faces of the query name and total number of images retrieved as in Table 4.4. The first column of the table refers to total number of true images retrieved vs. total number of true images retrieved by using only the speech transcripts -selecting the shots within interval [-1,2]. The numbers after removing the detected anchorpeople by the algorithm from the text-only results are given in the second column. And the last column is for applying the

algorithm to this set, from which the anchorpeople are removed. The precision values are given in Figure 4.14. Some sample images retrieved for each person are shown in Figure 4.13.

Table 4.4: Numbers in the table indicate the number of correct images retrieved/ total number of images retrieved for the query name.

Query name	Clinton	Netanyahu	Sam Donaldson	Saddam	Yeltsin
Text-only	160/2457	6/114	102/330	14/332	19/157
Anchor removed	150/1765	5/74	81/200	14/227	17/122
Method applied	109/1047	4/32	67/67	9/110	10/57

As can be seen from the results, we keep most of the correct faces (especially after anchorperson removal), and we get reject many of the incorrect faces. Hence the number of images presented to the user is decreased. Also, our improvement in precision values are relatively high. Average precision of only text based results increases by 29% after anchorperson removal, and by 152% after applying the proposed algorithm.

4.5 A Method for Finding the Graph Threshold Automaticaly

The normalized cut metric presented in [45] can be used for selecting the threshold to convert the weighted graph to a binary one. Let A be the set of vertices of a cut and V be the set of all vertices in graph G. Then the value of a cut and normalized cut of A are defined as follows:

$$cut(A,V) = \sum_{u \epsilon A, v \epsilon V - A} w(u,v),$$

$$Ncut(A,V) = \frac{cut(A,V)}{assoc(A,V)} + \frac{cut(A,V)}{assoc(V-A,V)},$$

where w(u, v) is a function of similarity between nodes u and v; and $assoc(A, V) = \sum_{u \epsilon A, t \epsilon V} w(u, t)$ is the total connection from nodes in A to all nodes in the graph.

We obtain different binary graphs by choosing different threshold in the range 0.55 and 0.65. Then graph algorithm is run on these graphs separately and the normalized cut value for each graph is calculated as defined as above. Among all, the thresholds applied in constructing the graph with minimum Ncut value is selected. The overall weighted recall and precision values are achieved as 74.01 and 68.55 respectively.

4.6 Performance Analysis

The performance of our system is mainly based on computing the similarity values since we compare each face with all other faces in the search space. Hence, the time complexity of constructing the similarity matrix is $O(N^2)$, where N is the total number of images in the limited search space of a query name. The time complexity of the greedy graph algorithm is $O(N)$; and it takes constant time to check if a new image belongs to the query person after the result of the graph approach is modeled.

To form an example, the similarity matrix of a search space with 200 pictures is constructed in 9 minutes on a Pentium IV 3 GHz machine with 2 GB memory; and it takes less than 1 second to partition this graph with the densest component algorithm.

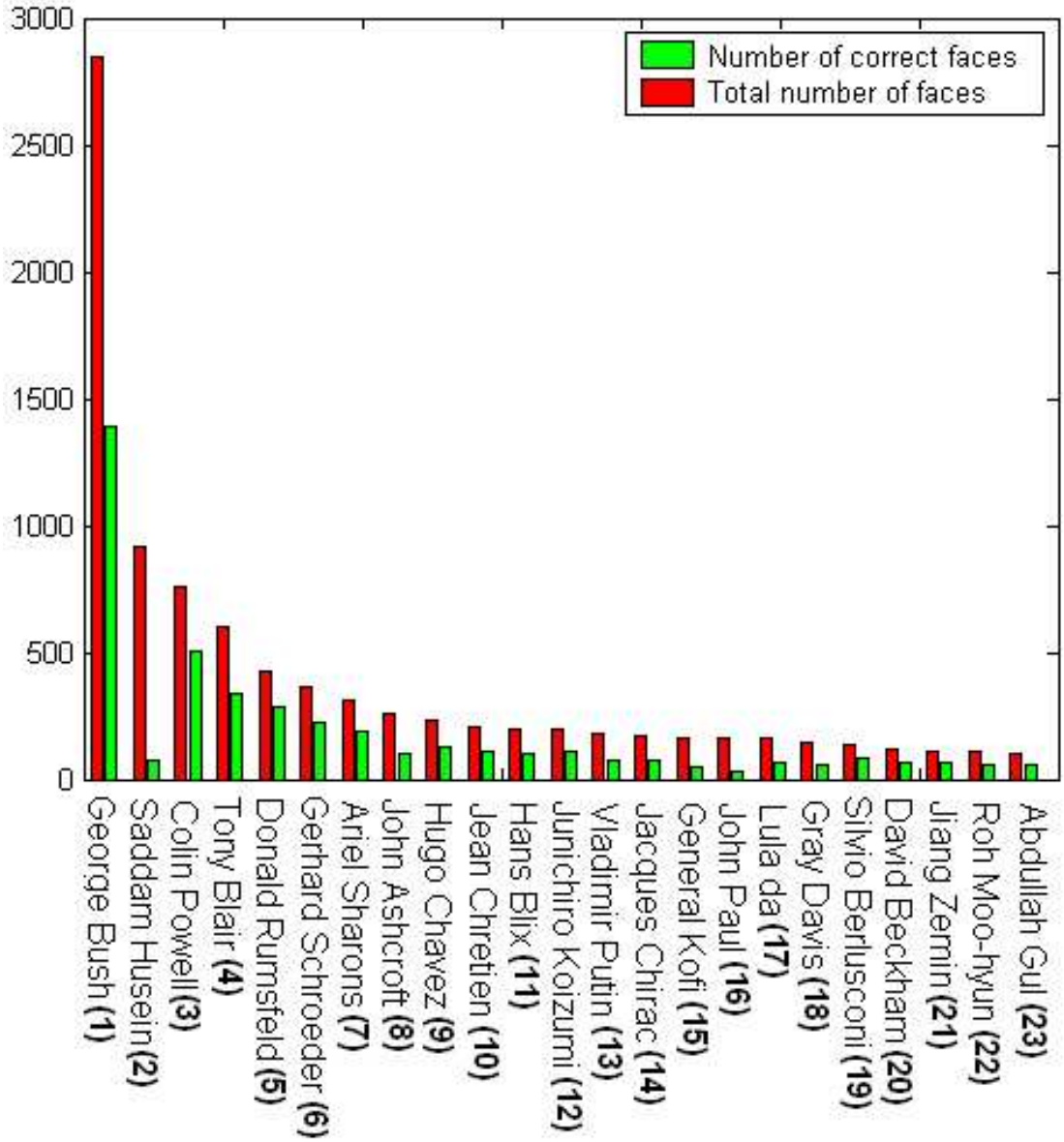


Figure 4.1: Names of 23 people are used in the experiments. The total number of faces associated with a name is represented by red bars and number of correct faces by green bars.

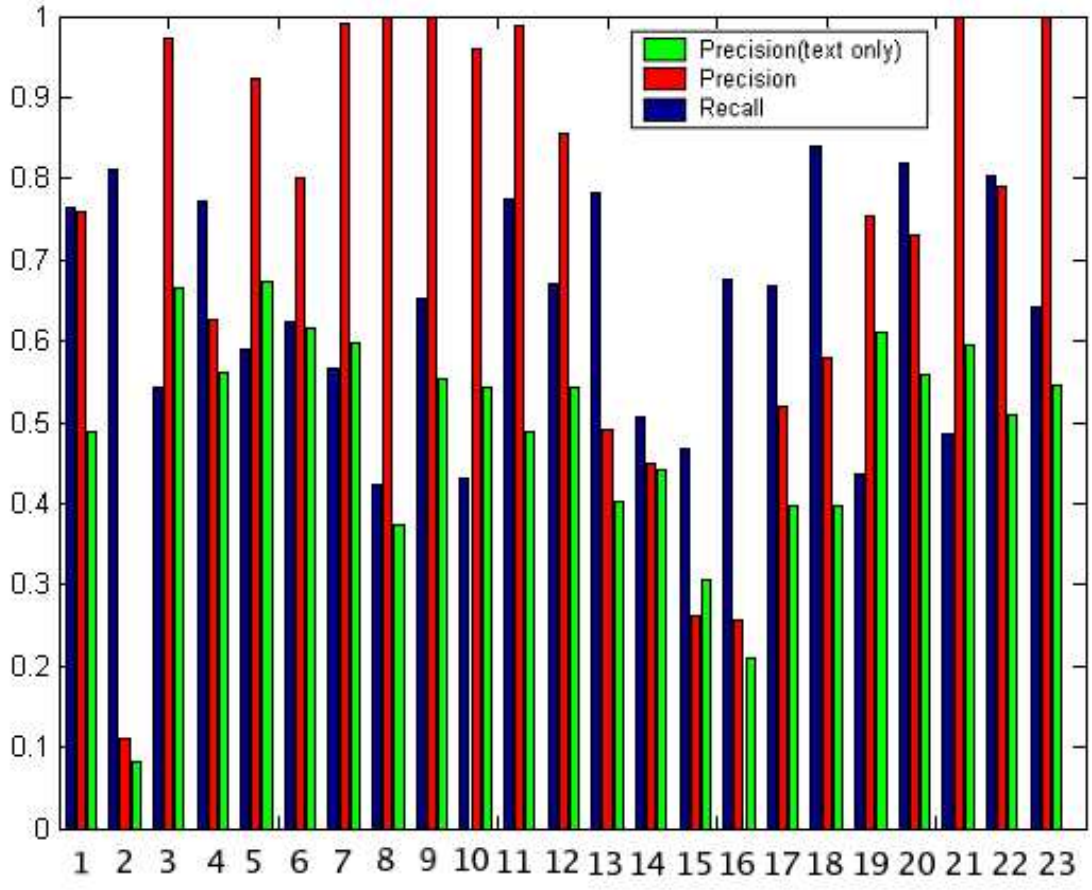
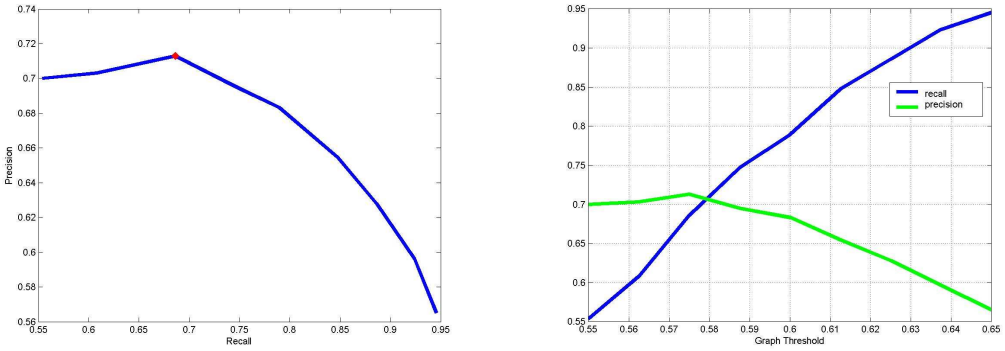


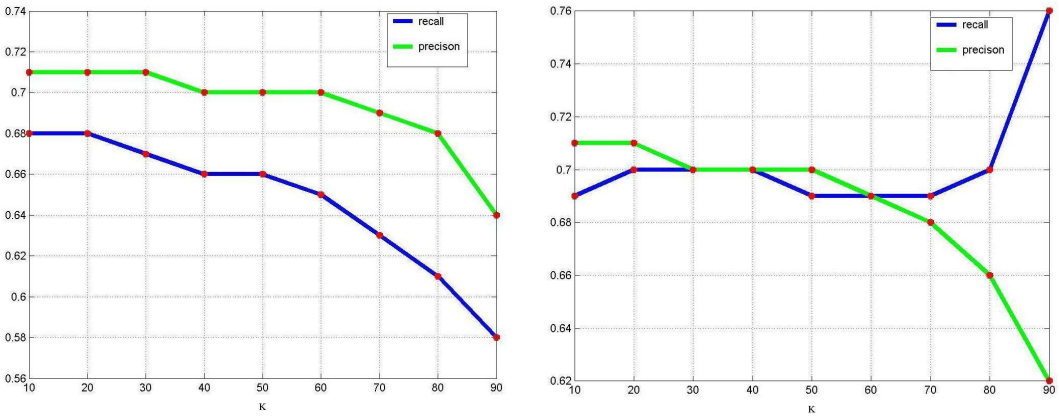
Figure 4.2: Recall and precision values for 23 people for graph threshold value of 0.575. Blue bars represent recall and red bars represent precision values that are achieved with the proposed method. Green bars are precision values for the baseline method, which does not use the visual information and retrieves the faces when name appears in the caption.



(a) (b)

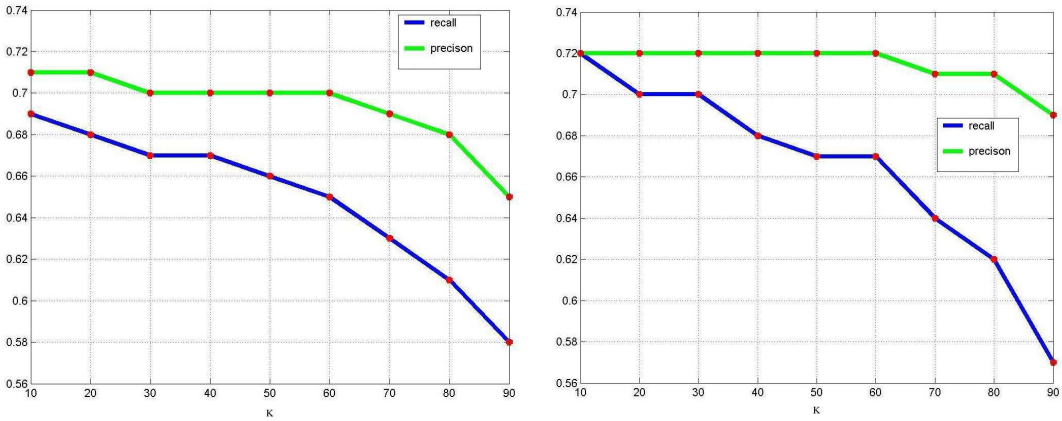
Figure 4.3: Weighted average recall and precision values of 23 people in the news photographs dataset. (a) Recall-precision curve depending on the graph threshold. The threshold used in the rest of our experiments is marked with red. (b) Recall and precision values as a function of the graph threshold.

Figure 4.4: Sample images retrieved (on the left) and sample images not retrieved (on the right) for the query names: George Bush (top), Colin Powell (middle), Hans Blix(bottom).



held-out set model

Figure 4.5: Recall and precision values of the held-out set (on the left) and the model (on the right) for online recognition with degree modeling.



held-out set model

Figure 4.6: Recall and precison values of the held-out set (on the left) and the model (on the right) for online recognition with distance modeling method.

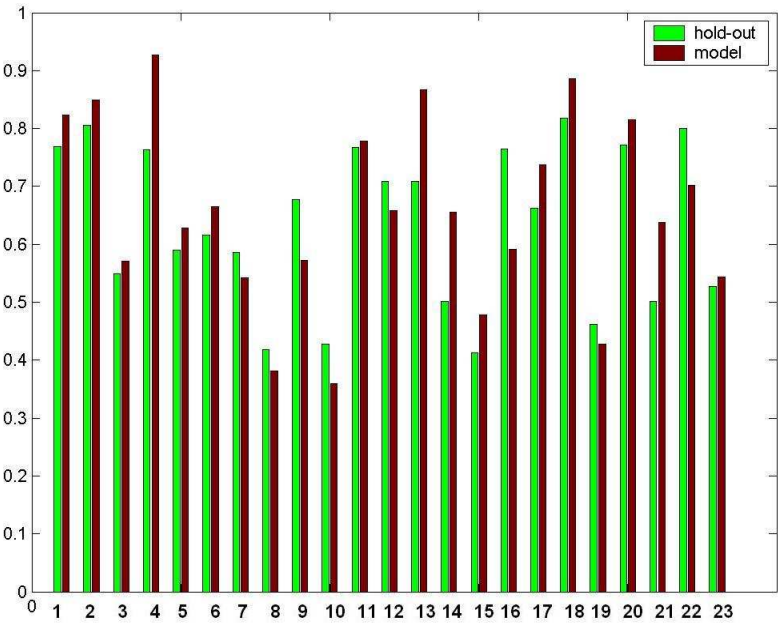


Figure 4.7: Recall values of the held-out set and the constructed model of each 23 people for K = 10.

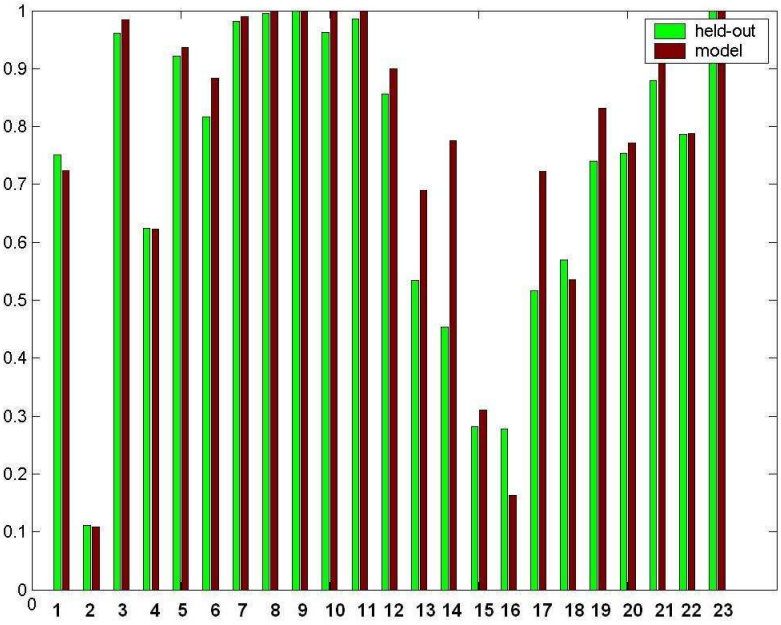


Figure 4.8: Precison values of the held-out set and the constructed model for each person for K = 10.

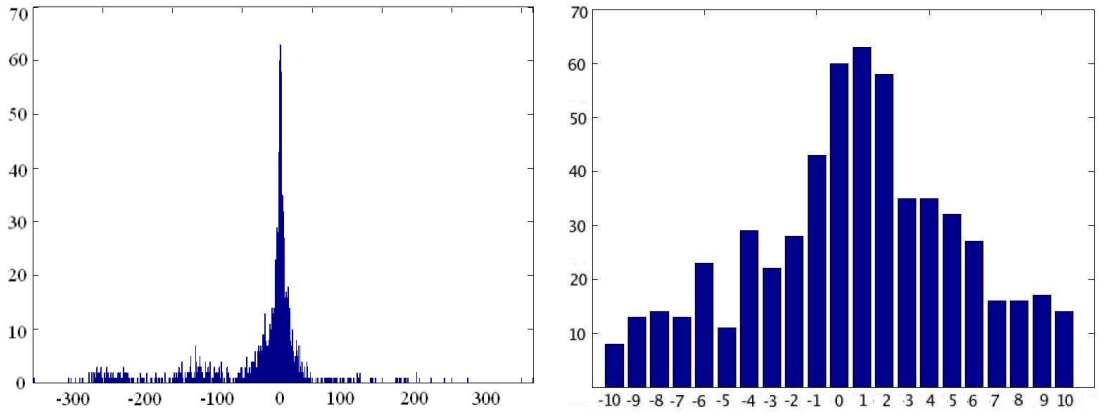


Figure 4.9: The figure shows frequency of Bill Clinton's visual appearance with respect to the distance to the shot in which his name is mentioned. **Left:** when the whole dataset is considered, **Right:** when the faces appearing around the name within the preceding and the following ten shots are considered. Over the whole dataset Clinton has 240 faces and 213 of them appear in the selected range.

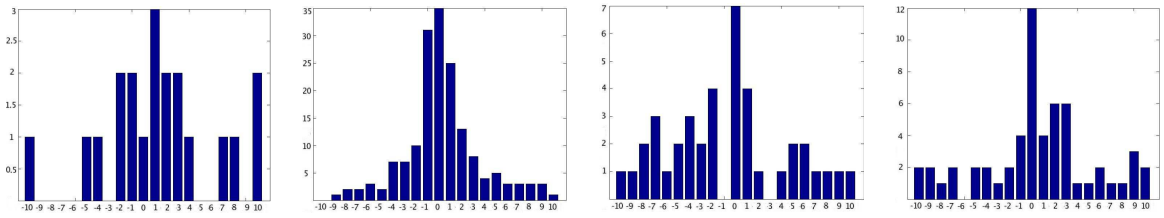


Figure 4.10: The relative position of the faces to the name for Benjamin Netanyahu, Sam Donaldson, Saddam Hussein, and Boris Yeltsin respectively.

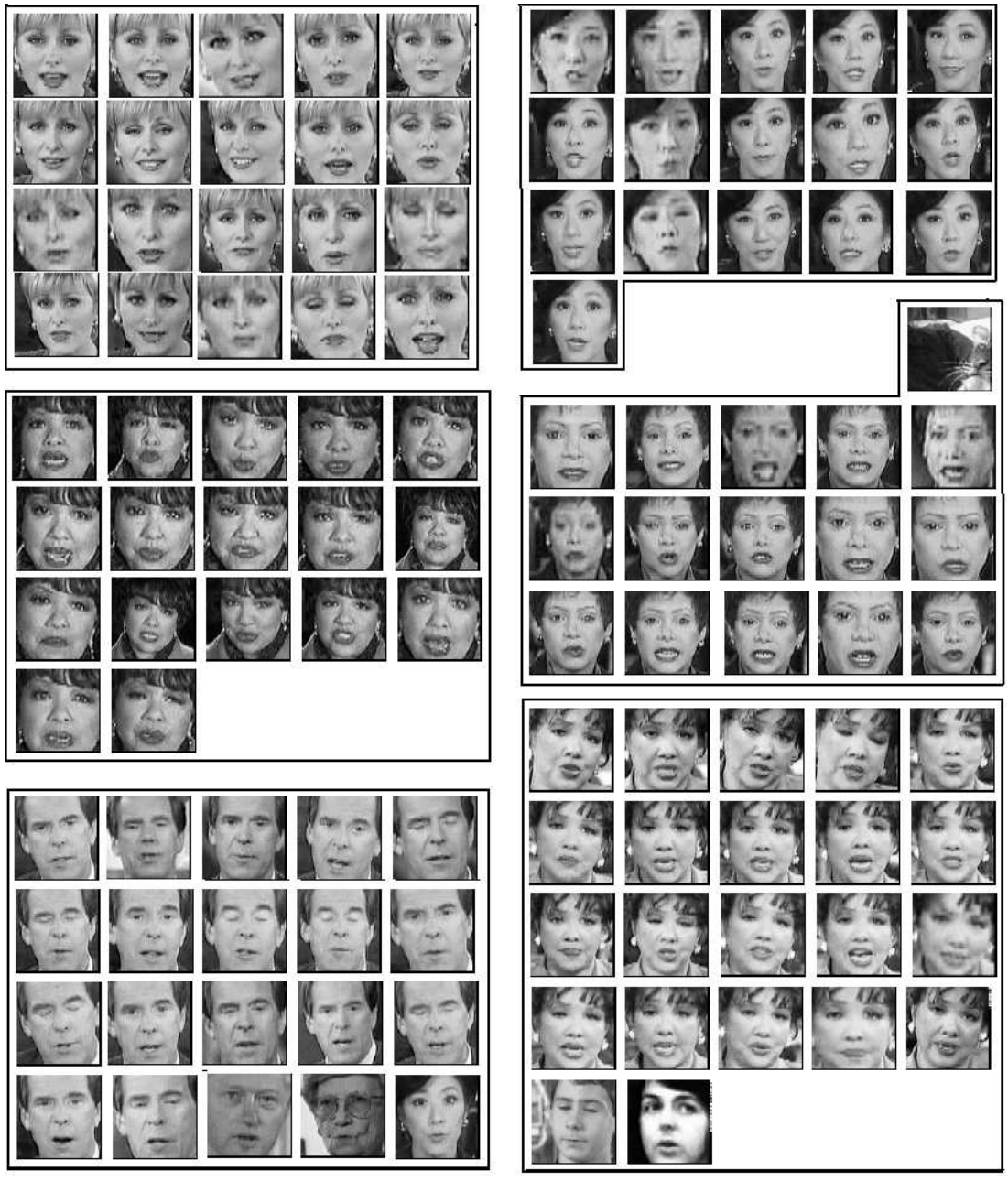
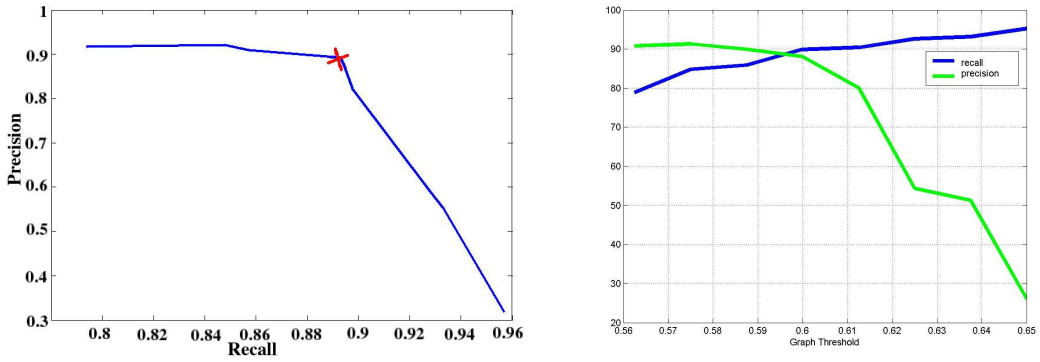


Figure 4.11: Detected anchors for 6 different videos.



(a) (b)

Figure 4.12: Weighted average recall-precision values of randomly selected 10 news videos. (a) Recall-precision curve depending on the graph threshold. The threshold used in the rest of our experiments is marked with red. (b) Recall and precision values as a function of the graph threshold.

Figure 4.13: Sample images retrieved for five person queries in experiments. Each row corresponds to samples for Clinton, Netanyahu, Sam Donaldson, Saddam, Yeltsin queries respectively.

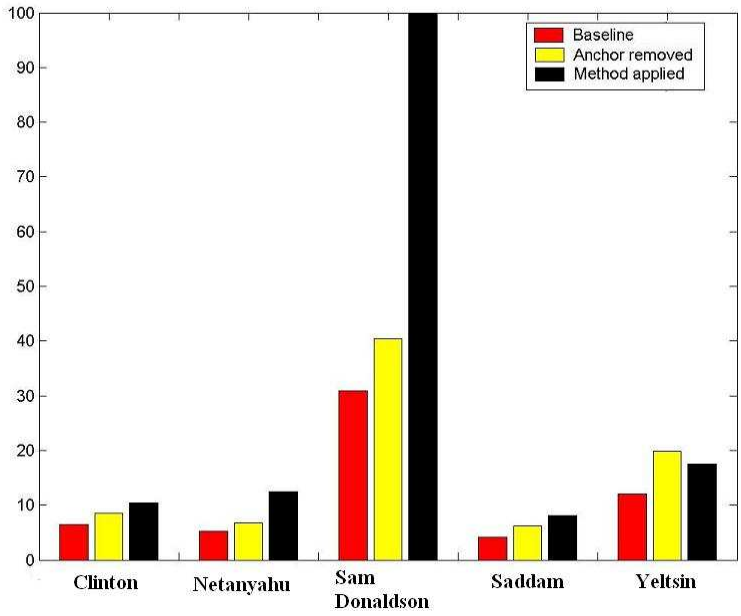


Figure 4.14: Precisions values achieved for five people used in our tests.

Chapter 5

Comparison

For comparison reasons, we first apply a traditional face recognition algrithm on the same dataset as the baseline method. Then, we analyze two main parts of person finding approach separately: definition of the similarity measure in graph construction and the densest component algorithm for partitioning this graph. Finally, we compare our results with two other related studies in the literature.

5.1 Baseline Method

The principle component analysis (pca) is a well-known method that has also been used in face recognition as eigenfaces [35]. As to compare with a baseline method, we have applied it on the news photographs data, to give an idea of the performance of the traditional face recognition methods. The experiments are conducted on the ground truth faces of the top 23 people used in experiments. For each person, (100-K) percent of the images are selected for training. Remaining K percent of the images are then classified as being one of these people in the train set. The algorithm is run 10 times with different random groups of images for testing and training.

Recognition rate is calculated by dividing the truly labeled images by total number of images tested. Average recognition rates of both test and train set for different K are given in Table 5.1 and in Figure 5.1, which shows off relatively low rates for the test images and reaches only 0.52 for K = 10.

Table 5.1: Recognition rates of the eigenface method for for different K values.

K	10	20	30	40	50	60	70	80	90
test	0.52	0.52	0.50	0.49	0.46	0.43	0.40	0.35	0.28
train	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99

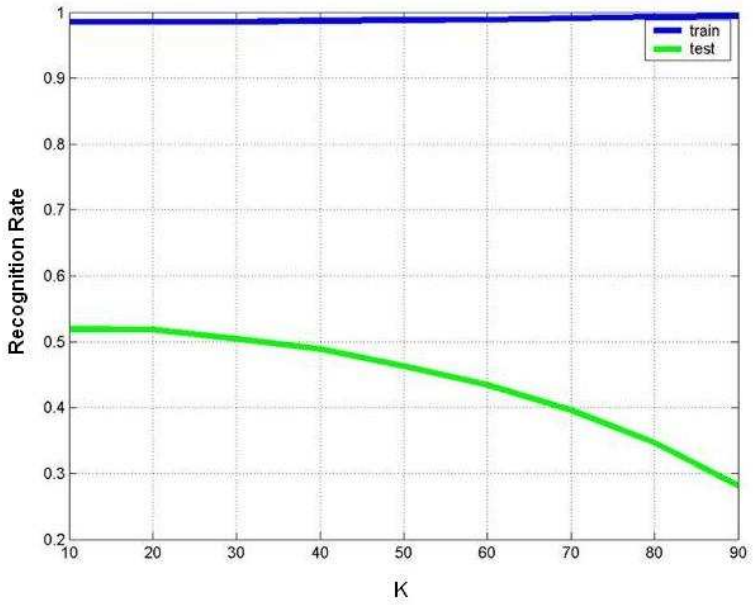


Figure 5.1: Recognition rates of the eigenface method for test and train sets as a function of K.

5.2 Feature Selection and Similarity Matrix Construction

5.2.1 Finding True Matching Points

As stated earlier, we have not used the original matching metric of SIFT, since it does not work well for faces. Some sample matching points found by using the original matching metric and the proposed method for the same images are shown in Figure 5.2. As the results indicate, using the original metric misses many possible matches and we can keep those matches with the proposed scheme.

To see how well the original metric performs, we have constructed the similarity graph by using the matches of the original metric. Then, we have applied the densest component graph algorithm on this similarity graph. The performance results for different graph thresholds are plotted in Figure 5.3. The recall and precision values for the same threshold in our original tests (0.575) are recorded as 0.91 and 0.59 respectively, which were 0.68 recall and 0.71 with the proposed approach. Although the average recall values seems to be relatively high, average precision does not get through the 0.61 value.

In a second experiment, we have applied the Ransac algorithm [19] on our initial matching points obtained before applying the two constraints. By this way, we have found the tomography matrix and assigned true matches using it. Sample results of this techniques are given in Figure 5.4. As it can be perceived from the results, affine constraints does not work well due to deformability property of the faces.

5.2.2 Facial Features

Many of the face recognition methods use facial features, which refer to eyes, mouth, nose and middle region of the two eyes. For comparison, we have manually labeled those regions for 5 people from news photographs dataset (Hans Blix,

Jacques Chirac Kofi Annan, Gray Davis, and David Beckham). A sample for selected facial regions is shown in Figure 5.5.

We have represented each facial feature with a SIFT descriptor and found the similarity matrices by comparing those feature descriptors. Average recall and precision values are given in Figure 5.6 for both the facial feature approach and the proposed approach. For all those 5 people, the proposed method achieves better precision values. Although the recall values are high for facial features approach, the precision values tend to be more close to the baseline (text-only precision).

5.3 Extracting Similar Group of Faces

We compare the greedy graph algorithm for finding the densest component with one well-known approach: k-nearest neighbor and the one-class classification methods in the following two subsections. All the experiments are conducted on news photographs dataset, where we had achieved 0.68 recall and 0.71 precision values on the average.

5.3.1 k-nn Approach

In these experiments a method similar to the k-nearest neighbors classification has been used. For each face in the test set, we find the distances of that face to all the faces in the training set, and select the nearest k faces (k-neighbors). If number of true faces are greater than the number of false ones in this k-neighbors, then the test face is classified as a true face. The tests are conducted with different number of training and testing sets. Each time, we have selected P percent of the images in a search space randomly for testing, and the rest for training. The algorithm is then run 10 times for each P.

Average results are given in Table 5.2 and in Figure 5.7. The highest recall and precision values are 0.68 and 0.41 respectively when k = 3, Average precision value of the baseline (text-only results) decreases from 0.48 to 0.41 with this method. Hence, the results indicate that the greedy densest component algorithm outperforms the k-nn approach.

Table 5.2: Recognition rates of supervised method for different P values. (K is the percentage of the images that are used in testing; k is the number of neighbours used.

P	10	20	30	40	50	60	70	80	90
k(neighbours) = 11									
recall	0.53	0.52	0.51	0.49	0.47	0.45	0.41	0.36	0.33
precision	0.36	0.36	0.35	0.34	0.34	0.33	0.32	0.31	0.31
k(neighbours) = 5									
recall	0.61	0.61	0.60	0.58	0.56	0.54	0.51	0.47	0.40
precision	0.38	0.38	0.38	0.38	0.37	0.37	0.36	0.35	0.34
k(neighbours) = 3									
recall	0.68	0.67	0.65	0.64	0.62	0.59	0.56	0.52	0.45
precision	0.41	0.40	0.40	0.40	0.39	0.39	0.38	0.37	0.36

5.3.2 One-class Classification

Given a set of data items, one-class classification methods aim to find a target class against the outliers([48]). In other words, given a test sample, it is either accepted as belonging to the target class or rejected. Hence, one-class classification approach differs from any other traditional multi or two-class classification approaches by holding only the information of the target class. With the greedy densest component algorithm [13], we also seek to find only the nodes belonging to the densest component (hence the faces of the query person) and assume all others are outliers. In this context, the most similar problem to the problem of finding the densest component of the graph is one-class classification problem. To this end, we compare the one-class classification methods in [48] with the graph algorithm used in this study.

The similarity graph constructed as described in 3.3 keeps only the distances among faces. Hence, the one-class classification methods cannot be applied to this graph. So to compare the greedy graph densest component method with any of those methods, we used the Bag-Of-Features approach as in [46, 32] for graph construction. Then, we applied both the greedy graph method and two of the one-class classification methods (nearest neighbor data description method and k-nearest neighbor data description method) that gave us the best results among all.

To construct the graph, we first extracted sift features from each face image and clustered these features using k-means clustering into 50 clusters. Then, a histogram of the size of number of clusters (50) is formed for each image showing the distribution of clusters. In Information Retrieval, the frequencies of the clusters are weighted by 'term frequency inverse document frequency (tf-idf)' which is computed as:

$$tfidf = \frac{n_{id}}{n_d} \log \frac{N}{n_i} \tag{5.1}$$

where, n_{id} is number of occurrences of term i in document d, n_d is total number of terms in document d, N is the total number of documents in database and n_i is the number of documents in database containing term i.

Adapting the same approach, we find the weighted frequencies of the clusters and use them as the final feature vector for each image. The one-class classification methods can then be applied on these features. To apply the densest component graph algorithm, the similarity among the faces are found by normalized scalar product of tfidf vectors by using the following equation:

$$distance(f_1, f_2) = \frac{tfidf(f_1) * tfidf(f_2)}{norm(f_1) * norm(f_2)} \tag{5.2}$$

where, $tfidf(f_1)$ is tf-idf vector of face image f_1 and $norm(f_1)$ is the norm of tfidf vector of face image f_1.

Table 5.3: Recall-precision rates of two one class classification methods: w1 (nearest neighbor data description method) and w2 (k-nearest neighbor data description method) (applied on tfidf's).

training set										
	K = 10		K = 20		K = 30		K = 40		K = 50	
	rec	pre	rec	pre	rec	pre	rec	pre	rec	pre
w1	1.00	0.54	1.00	0.53	1.00	0.52	1.00	0.52	1.00	0.51
w2	0.90	0.57	0.90	0.55	0.90	0.54	0.90	0.54	0.90	0.53
test set										
	K = 10		K = 20		K = 30		K = 40		K = 50	
	rec	pre	rec	pre	rec	pre	rec	pre	rec	pre
w1	.90	0.50	0.91	0.50	0.90	0.50	0.90	0.50	0.90	0.49
w2	0.84	0.53	0.88	0.54	0.87	0.53	0.86	0.53	0.86	0.52

The precision-recall curve of the densest component graph algorithm for varying graph thresholds is given in Figure 5.8. Recall and precision values of one-class-classification methods for different K-fold validation tests are given in Table 5.3. For the first one-class classification method, nearest neighbor data description method, average precision is around 0.50 where average recall goes around 0.90. Similarly for the second method, k-nearest neighbor data description method, average precision is around 0.86 and average recall around 0.53. For a recall value of around 0.86 and 0.90, we also achieve similar precision values as with the one-class classification methods. This indicates that the once-class classification approach is not superior to the greedy densest component algorithm used in our experiments for finding the most similar set of faces.

5.4 Comparison with Related Studies

In this section, we compare the proposed method with two related studies in literature.

The first study has been presented by Berg et al in [8]. As stated previously in Section 2.1, a method for associating faces with a set of names is proposed in that study. The method is tested on the same dataset that we have used in news photographs experiments.

We have plotted in Figure 5.9 and 5.10, the recall and precision values of the same 23 people as in our experiments. The related study achieves better recall values. However, we also get better precision values for many people. Precision values of the baselie method, precision in the limited search space of a query perosn, are also given in 5.10. The cases where we exceed the previous study is mostly when the baseline precision is relatively high. In other words, we perform better when if our initial assumption that a person appears the most (than others) and forms that largest similar group of images in its limited search space holds.

The second study, which has been presented in [41], is a system for re-ranking the results of a search engine by exploiting from both keyword and content based retrieval. The system downloads and works on the first 500 images from Yahoo Image API for a given query. First, it segments each image into blobs and builds a color histogram for each blob. Then, it clusters the blobs via mean-shift clustering [16] and finds the cluster of blobs corresponding to the largest number of parent images. This cluster is called the significant cluster. Hence, its mean is then used for re-ranking all the search results based on distance of each blob in each image to the significant cluster.

The idea of significant cluster presented in that study is similar to the idea of largest densest component presented in this thesis. If we consider each face image in a limited search space of a query person as one blob, then the largest densest component of this space refers to the significant cluster.

For comparison, we first applied multi-dimensional scaling [31] on the weighted similarity matrix of each 23 people in news photographs dataset and found the x-y coordinates of each item (face images) in a matrix. Using their x-y coordinates, we clustered these images also by the mean-shift clustering method and found the significant cluster corresponding to the faces of the query person. Recall and precision values depend on the window size used in mean-shift clustering. Thus,

we run the algorithm for varying window sizes and plotted the recall-precision values as a function of the window size in Figure 5.11. Although the recall value can increase up to 1.00, the maximum precision value reported is 0.51, which is close to text-only results (0.48). Hence, it does not improve the text-only results.

In a second experiment, we have applied the significant clusters by applying the mean-shift clustering method on the features obtained by the Bag-Of-Features approach. (Details of the Bag-Of-Features approach is explained in the previous section). Results are given in Figure 5.12. Similarly, although the recall value is around 1.00, precision value remains around baseline (0.48); hence has no improvement on the text-only results.

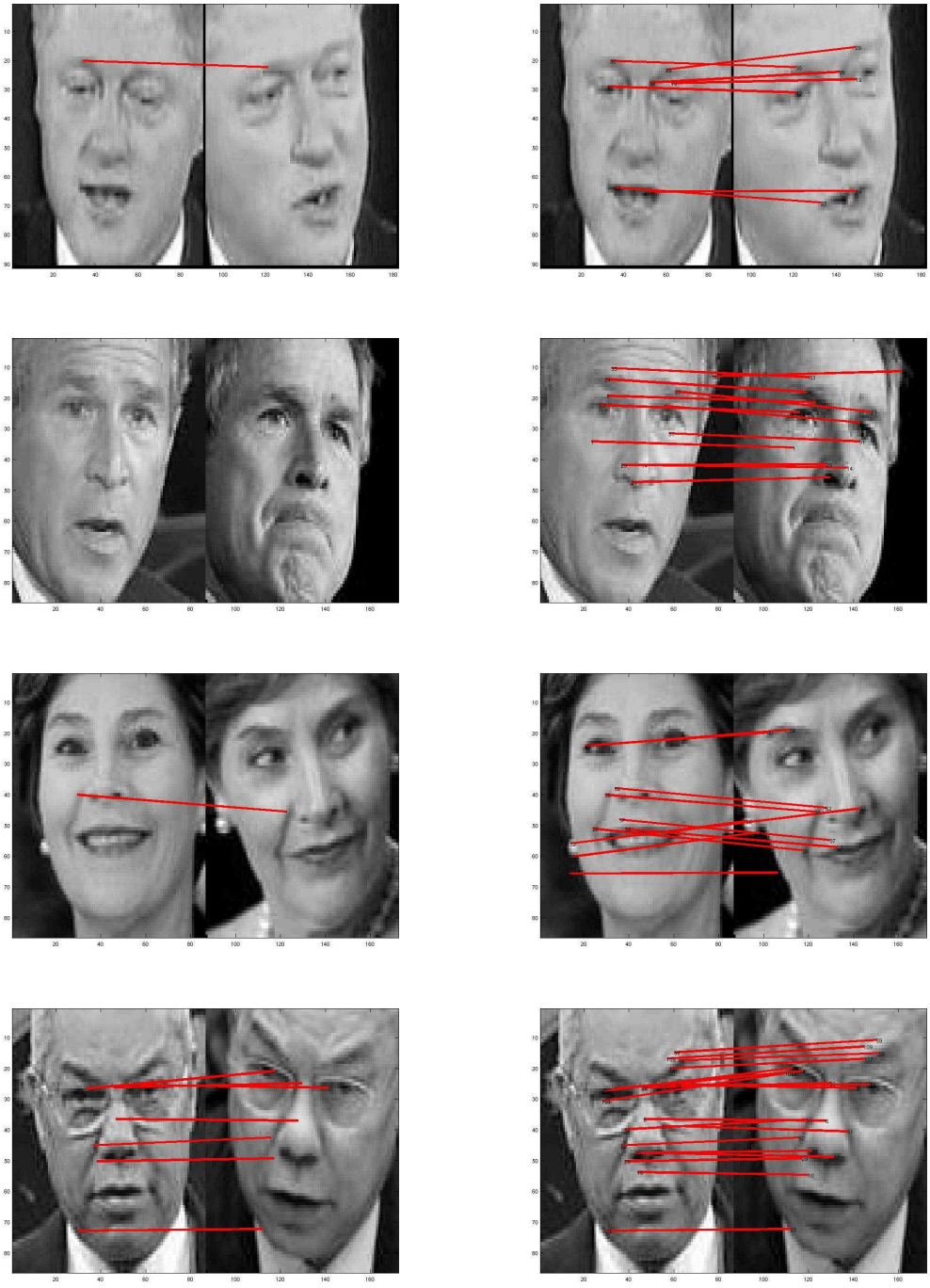
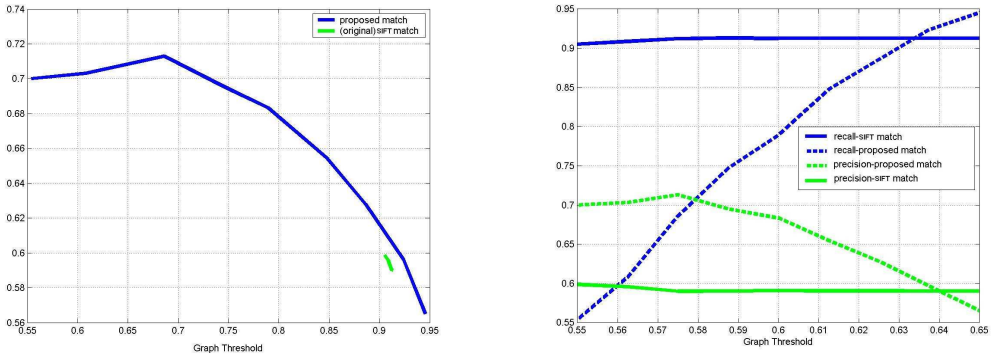


Figure 5.2: Examples for matching points. First column for matches found by using the original matching metric of sift. Second column is for matches found by applying the proposed method.



(a) (b)

Figure 5.3: Average recall-precision values of 23 people in the news photographs dataset obtained by using original SIFT matching metric for graph construction. For comparison, we also plot the values gained by the proposed method. (a) Recall-precision curve depending on the graph threshold. (b) Recall-precision values as a function of the graph threshold.

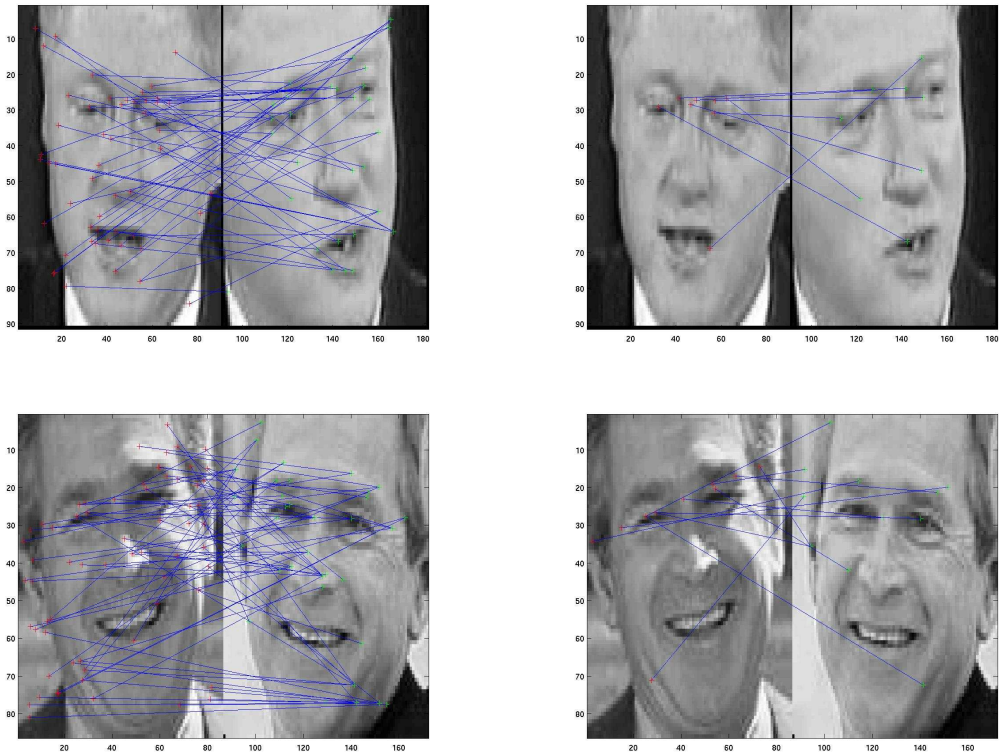


Figure 5.4: Examples for matching points assigned by the homography matrix found after applying the ransac algorithm.

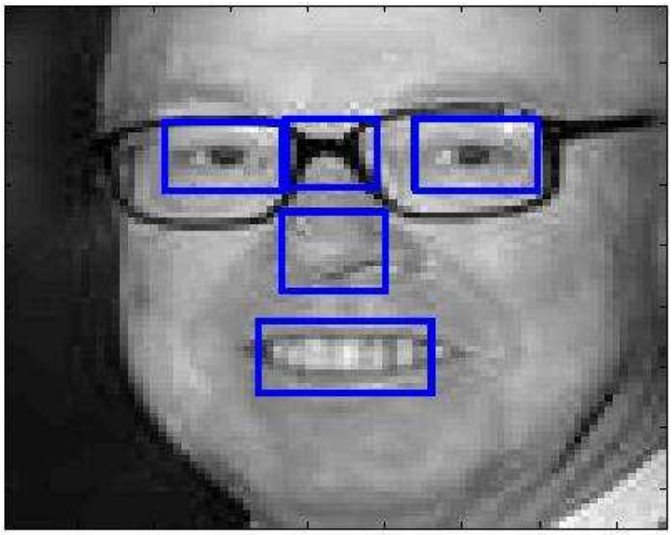
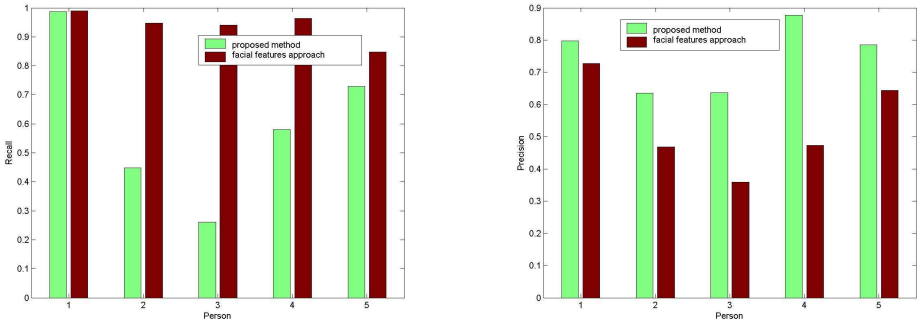
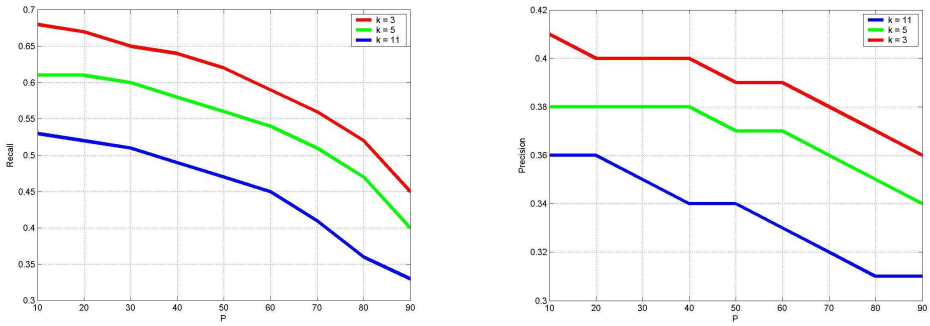


Figure 5.5: A sample for selected facial regions.



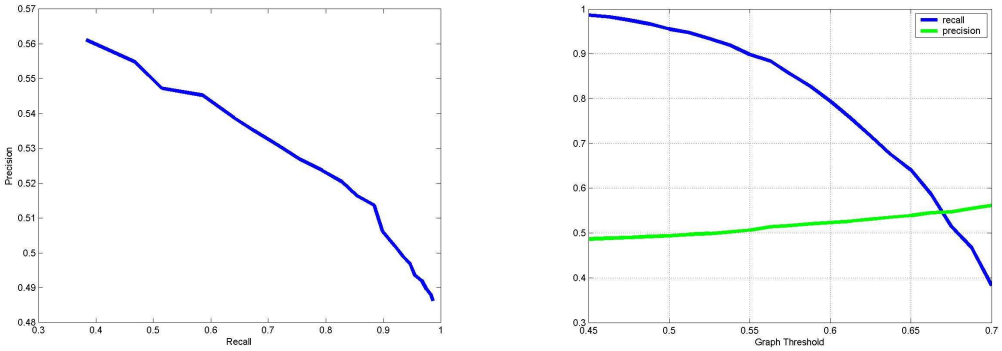
recall precision

Figure 5.6: Recall and precision values of 5 people from news photographs dataset when only the SIFT features around the facial features used.



recall precision

Figure 5.7: Recognition rates of the k-nn approach for different P and k values. P is the percentage of the images used for testing, and k is the number of neighbors in k-nn.



(a) (b)

Figure 5.8: Recall and precision values of 23 people in the test set when the greedy densest component algorithm is applied on the features obtained with the Bag-Of-Features approach. (a) Recall-precision curve depending on the graph threshold. (b) Recall-precision values as a function of the graph threshold.

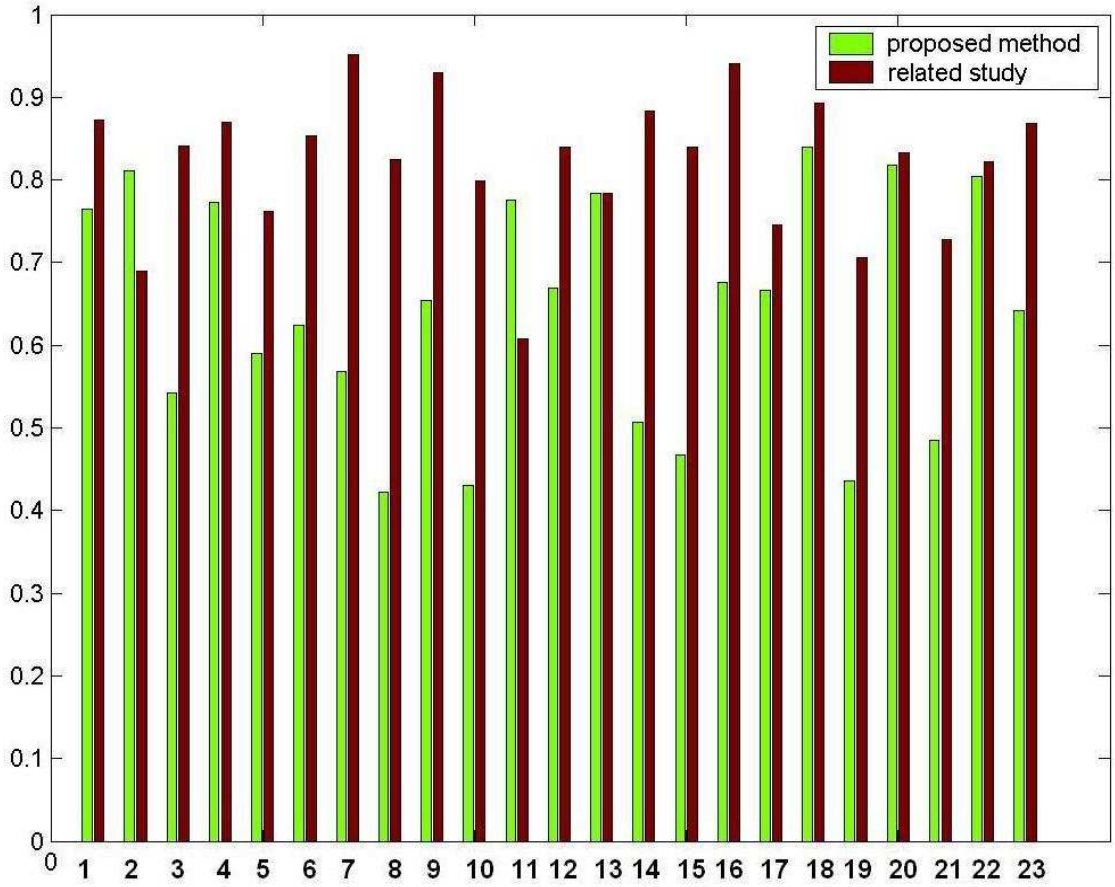


Figure 5.9: Recall values of the related study and the proposed method for each 23 people from news photographs dataset.

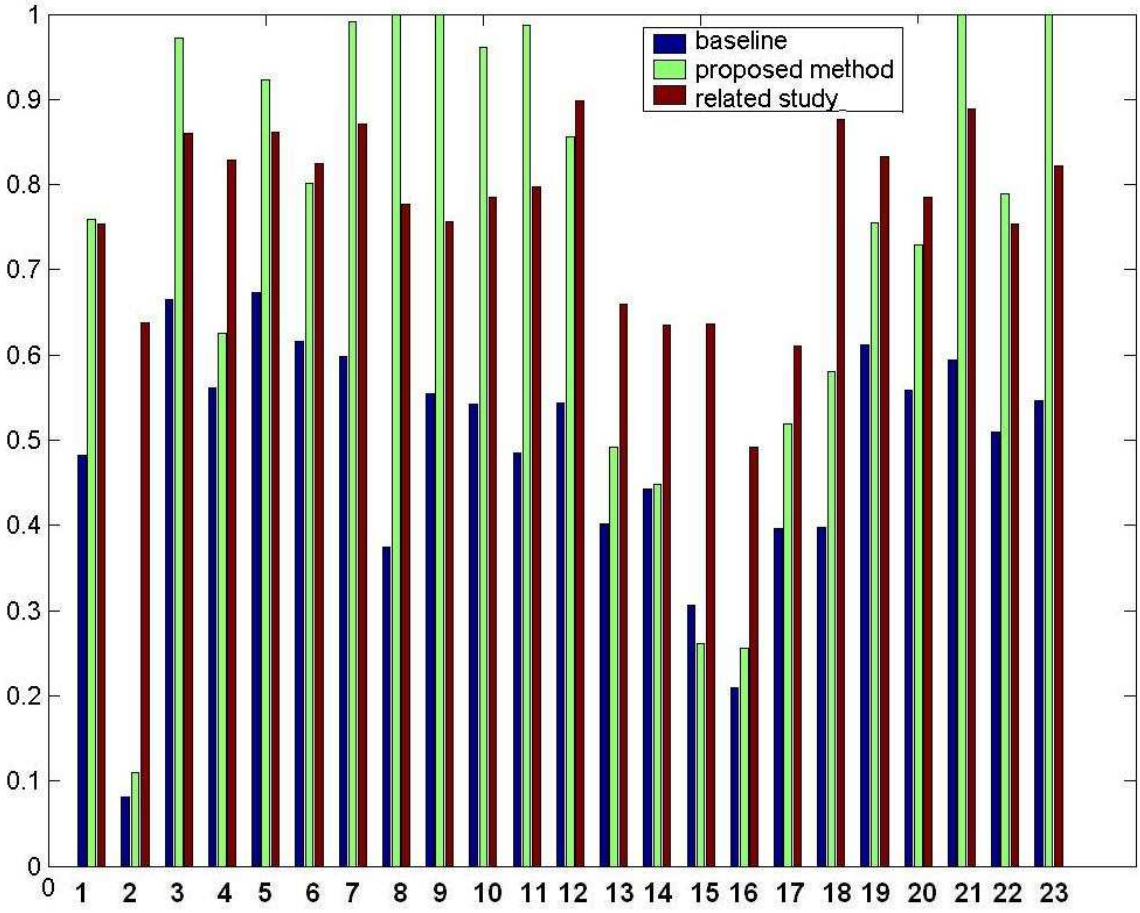
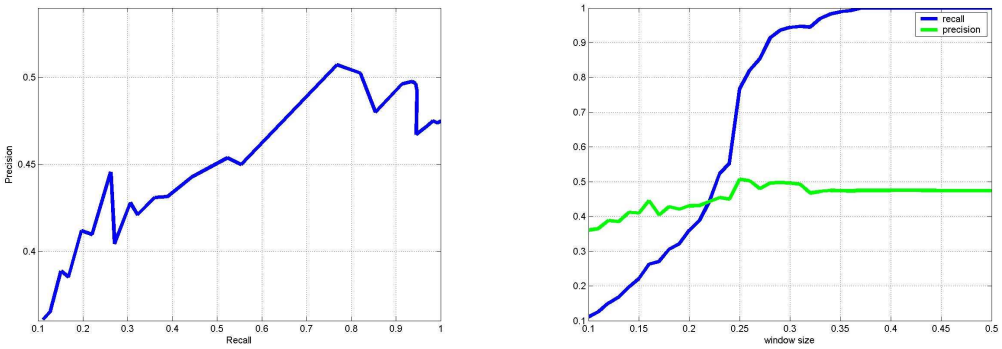
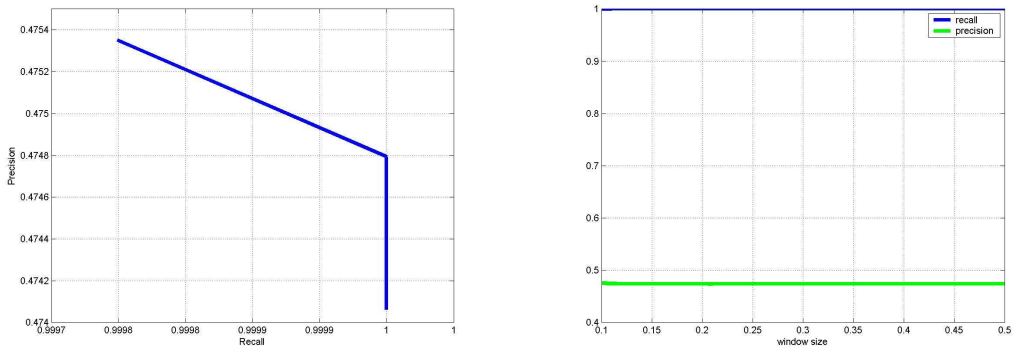


Figure 5.10: Precision values of the related study and the proposed method for each 23 people from news photographs dataset.



(a) (b)

Figure 5.11: Recall and precision values for 23 people in the news photographs dataset, when the significant cluster approach is applied on multi-dimensional scaling coordinate features. (a) Recall-precision curve depending on the window size. (b) Recall-precision values as a function of the window size used in mean-shift clustering.



(a) (b)

Figure 5.12: Recall and precision values for 23 people in the news photographs dataset, when the significant cluster approach is applied on Bag-Of-Features. (a) Recall-precision curve depending on the window size. (b) Recall-precision values as a function of the window size used in mean-shift clustering.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we propose a graph based method for querying people in large news photograph and video collections with associated captions or speech transcript texts. Given similarity measures between the face images in a data set, the problem is transformed into a graph problem in which we seek the largest densest component of the graph corresponding to the largest group of similar faces. We use SIFT descriptors [34] to represent each face image and define the similarity values by using the average distances of the matching interest points. Then, we apply a greedy graph algorithm [13] to find the densest component of the graph corresponding to the faces of the query person. In the study, we also propose two different methods to use the results of the graph based person finding approach in futher recognition of new faces.

For large realistic data sets, face recognition and retrieval is still a difficult and an error-prone problem due to large variations in pose, illumination and expressions. In this study, we have described a multi-modal approach for querying large numbers of people in such data sets. The method does not require training for any specific person and thus it can be applied to any number of people. With this property, it is superior to any supervised method which requires labeling of

large number of samples. The results achieved are also very close to supervised methods.

The experiments are conducted on two different news data sets. The first set consists of thousands of news photographs with associated captions collected from Yahoo! news. The captions are used to limit the number of images for a query name and only the images associated with the name are selected. In this data, over 20% increase in precision is achieved compared to solely text-based methods. For individuals, up to 84% recall and 100% precision values can be obtained.

The second experiments are conducted on 229 broadcast news videos archive. We fist use the speech transcripts and select the neighboring shots in which the name of the query name appears to limit our search space. Applying the proposed person finding algorithm on each video separately, we detect the anchorperson in each video. Then, we remove detected anchorperson from the search space of the query name and apply the algorithm to the remaining images. Experiments show that we improve person search performances relative to only text based results. Average precision values of only text based results are increased by 29% after anchorperson removal, and by 152% after applying the proposed algorithm. The person finding algorithm also performs well for anchorperson detection without requiring any supervision.

The proposed method is an overall scheme to find and recognize faces in large news photograph and video collections. Each step of the method can be commited with another tecniques, for instance similarities between faces can be assigned with a different approach or the densest component can be extracted with another graph partitioning algorithm. However, experiments show that our similarity definition works better than other traditional approaches for this dataset; and the greedy graph algorithm is comparable to one other the most possible approach, namely one-class classification.

One of the important remarks to be made on the method is that even if it is not a face recognition scheme on the whole, it is instrumental in reducing the number of images presented to the user by improving the retrieval performance

of baseline methods.

6.2 Future Work

Before applying the greedy densest component algorithm, we convert our weighed graph consisting of dissimilarity values into a binary graph. However, this ignores some of the information. A method, which does not violate the weighted property of the graph, may yield better results. In this study, SIFT descriptors are used to represent the similarity of the faces. Other representations or similarity measures can also be used to construct the graph structure.

In [26] sets of face exemplars for each person are gathered automatically in shots for tracking in video. A similar approach can be adapted and instead of taking a single face from each shot by only considering the key-frames, face detection can be applied to all frames to obtain more instances of the same. This approach can help to find better matching interest points and more examples that can be used in the graph algorithm.

Since the proposed approach is an overall scheme, it can also be applied to other problems such as object recognition or image region annotation. For instance, in the context of region annotation, annotations of images can be used for limiting the search space for a region. Similary, in this limited space, the region of interest is expected to form the largest similar group of regions.

Bibliography

[1] Trec video retrieval evaluation
http://www-nlpir.nist.gov/projects/trecvid/, 2004.

[2] J. Sivic . Everingham and A. Zisserman. Hello! my name is... buffy – automatic naming of characters in tv video. In *Proceedings of the British Machine Vision Conference*, 2006.

[3] S. Aksoy and R. M. Haralick. Graph-theoretic clustering for image grouping and retrieval. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 63–69, 1999.

[4] M. Bartlett, H. Lades, and T. Sejnowski. Independent component representations for face recognition, 1998.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts, 2001.

[6] S. Beretti, A. Del Bimbo, and E. Vicario. Efficient matching and indexing of graph models in content-based retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1089–1105, 2001.

[7] T. Berg, A. C. Berg, J. Edwards, and D.A. Forsyth. Who's in the picture. In *Neural Information Processing Systems (NIPS)*, 2004.

[8] T. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D.A. Forsyth. Names and faces in the news. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[9] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of sift features for face authentication. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 35, Washington, DC, USA, 2006. IEEE Computer Society.

[10] V. Bruce. Recognizing faces, 1988. Lawrence Erlbaum Associates, London, U.K.

[11] J.M. Fellous C. von der Malsburg, N. Kruger and L. Wiskott. Face recognition by elastic bunch graph matching. In *Computer Analysis of Images and Patterns 1997*, pages 456–463, 1997.

[12] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara CA USA, June 1998.

[13] M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In *APPROX '00: Proc. of the 3rd International Workshop on Approximation Algorithms for Combinatorial Optimization*, London, UK, 2000.

[14] F. Chen, U. Gargi, L. Niles, and H. Schuetze. Multi-modal browsing of images in web documents. In *Proceedings of SPIE Document Recognition and Retrieval VI*, 1999.

[15] M-Y. Chen and A. Hauptmann. Searching for a specific person in broadcast news video. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Montreal, Canada, May 17-21 2004.

[16] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[17] S. Dickinson, M. Pelillo, and R. Zabih. Introduction to the special section on graph algorithms in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1049–1052, 2001.

[18] P. Duygulu and A. Hauptmann. What's news, what's not? associating news videos with words. In *The 3rd International Conference on Image and Video Retrieval (CIVR) Ireland*, July 21-23, 2004.

[19] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[20] J.L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2), 2002.

[21] Y. Gdalyahu, D. Weinshall, and M. Werman. Self-organization in vision: Stochastic clustering for image segmentation, perceptual grouping, and image database organization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1053–1074, 2001.

[22] R. Gross, S. Baker, I. Matthews, and T. Kanade. Face recognition across pose and illumination. In Stan Z. Li and Anil K. Jain, editors, *Handbook of Face Recognition*. Springer Verlag, 2004.

[23] R. Gross, J. Shi, and J. Cohn. Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*, 2001.

[24] G. Hamerly and C. Elkan. Learning the k in k-means. In *Advances in Neural Information Processing Systems*, volume 17, 2003.

[25] S. Helmer and D.G. Lowe. Object recognition with many local features. In *Workshop on Generative Model Based Vision 2004(GMBV)*, Washington D.C., 2004.

[26] M. Everingham J. Sivic and A. Zisserman. Person spotting: video shot retrieval for face sets. In *International Conference on Image and Video Retrieval (CIVR), Singapore*, 2005.

[27] G. M. Davies J. W. Shepherd and H. D. Ellis. Studies of cue saliency, 1981. In Perceiving and Remembering Faces, G. M. Davies, H. D. Ellis, and J. W. Shepherd, Eds. Academic Press, London, U.K.

[28] I. Jermyn and H. Ishikawa. Globally optimal regions and boundaries as minimum ratio weight cycles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1075–1088, 2001.

[29] Y. Kaya and K. Kohayashi. A basic study on human face recognition, 1972. Frontiers of Pattern Recognition, S. Watanabe, ed., p. 265.

[30] Michael David Kelly. *Visual identification of people by computer*. PhD thesis, 1971.

[31] J.B. Kruskal and M. Wish. Multidimensional scaling, 1978. Sage University Paper Series on Quantitative Application in the Social Sciences. Sage Publications, Beverly Hills and London.

[32] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[33] H. Le and H. Li. Recognizing frontal face images using hidden markov models with one training image per person. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1*, pages 318–321, Washington, DC, USA, 2004. IEEE Computer Society.

[34] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 2004.

[35] A. Pentland M. Turk. Eigenfaces for recognition. *Journal of Cognitive Neurosicence*, 3(1):71–86, 1991.

[36] A.P. Pentland M.A. Turk. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1991.

[37] B. S. Manjunath, R. Chellappa, and C. von der Malsburg. A feature based approach to face recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1992.

[38] K. Mikolajczk and C. Schmid. A performance evaluation of local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[39] K. Mikolajczyk. Face detector. INRIA Rhone-Alpes, 2004. Ph.D Report.

[40] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision (ICCV)*, pages 525–531, 2001.

[41] B. Babenko N. B. Haim and S. Belongie. Improving web-based image search via content based clustering. In *Semantic Learning Applications in Multimedia 2006*, page 106, 2006.

[42] B. Park. Face recognition using face-arg matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1982–1988, 2005. Member-Kyoung-Mu Lee and Member-Sang-Uk Lee.

[43] P.E. Hart R.O. Duda and D.G. Stork. In *Pattern classification*. John Wiley and Sons, 2001.

[44] S. Satoh and T. Kanade. Name-it: Association of face and name in video. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 1997.

[45] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[46] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1470, Washington, DC, USA, 2003. IEEE Computer Society.

[47] X. Tan, S. Chen, Z. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition.*, 39(9):1725–1745, 2006.

[48] D.M.J. Tax. *One-class classification*. PhD thesis, Delft University of Technology, http://ict.ewi.tudelft.nl/ davidt/thesis.pdf, June 2001.

[49] R. Chellappa W. Zhao. Image-based face recognition: Issues and methods. *Image Recognition and Classification*, pages 375–402, 2002.

[50] J. Wang, K. N. Plataniotis, and A. N. Venetsanopoulos. Selecting discriminant eigenfaces for face recognition. *Pattern Recognition Letters*, 26(10):1470–1482, 2005.

[51] A. Webb. In *Statistical Pattern Recognition*. John Wiley and Sons, 2002.

[52] J. Yang, M-Y. Chen, and A. Hauptmann. Finding person x: Correlating names with visual appearances. In *International Conference on Image and Video Retrieval (CIVR)*, Dublin City University Ireland, 2004.

[53] J. Yang, D. Zhang, A. F. Frangi, and J. Yang. Two-dimensional pca: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):131–137, 2004.

[54] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

Appendix A

Different Forms of Names in News Photographs

George Bush	George W (1485), W. Bush (1462), George W. Bush (1454), President George W (1443), President Bush (905), U.S. President (722), President George Bush (44), President Bushs (2), President George W Bush (2), George W Bush (2)
Saddam Hussein	Saddam Hussein (911), President Saddam (356), President Saddam Hussein (351), President Saddam Hussien (3), President Saddam Hussiein (1), Minister Saddam (1)
Colin Powell	Colin Powell (618), Secretary of State Colin Powell (606) Secretary Colin Powell (4), Collin Powell (3), Secretary General Colin Powell (2), Secretary Powell (1), Secretary of State Powell (1)
Tony Blair	Tony Blair (502), Prime Minister Tony Blair (472), Premier Tony Blair (2), Prime Minister Blair (2), Mr Blair (1), Prime Minister Tony Bair (1)
Jean Chretien	Prime Minister Jean (158), Jean Chretien (155), Minister Jean Chretien (145), Prime Minister Jean Chretien (145), Prime Minister John Chretien (2), Prime Minister Chretien (2)
Gerhard Schroeder	Gerhard Schroeder (311), Chancellor Gerhard Schroeder (283) Chancellor Schroeder (1), Chancellor Gerhard Schroeders (2) Chancellor Gerhard Schroder (1), Chancellor Gerhard Schoeder (1)

John Ashcroft	John Ashcroft (147), General John Ashcroft (146) Attorney General John Ashcroft (143), U.S. Attorney (106) U. S. Attorney (2), U.S Attorney (1)
Donald Rumsfeld	Donald Rumsfeld (279), Donald H (47), Secretary Donald Rumsfeld (84), Donald H. Rumsfeld (44), Secretary Donald H (26), H. Donald (13), Secretary of State Donald Rumsfeld (4), Secretary Rumsfeld (6)
Ariel Sharons	Minister Ariel (249), Prime Minister Ariel (248) Prime Minister Ariel Sharons (2)
Junichiro Koizumi	Junichiro Koizumi (156), Prime Minister Junichiro (151) Prime Minister Junichiro Koizumi (149), Prime Minister Koizumi (1)
Hugo Chavez	Hugo Chavez (194), President Hugo Chavez (186), President Hugo Chaves (1), President Chavez (3)
General Kofi	General Kofi (124), Secretary General Kofi (61) Secretary-General Kofi (60), General General Kofi (1) Secretary-Genaral Kofi (1), Annan , U.N. (1), U.N. Secretary (57) U.N. Secretary- (39), U.N. General (9)
Roh Moo-hyun	Roh Moo-hyun (86), Roh Moo- (93), President Roh Moo-hyun (55) President-elect Roh Moo-hyun (10), President Roh (61), President Roh Moo- (61), President-Elect Roh Moo (1)
Lula da	Lula da (119), President Luiz Inacio Lula (30), President-elect Luiz Inacio Lula (19), President Lula (7), President Lula Da (5), President-elect Lula Da (2) President Luiz Lula Da (1), President Luis Inacio Lula (1) President-elect Luis Inacio Lula (1), Luiz Inacio (105), Luis Inacio (4)
Jacques Chirac	Jacques Chirac (143), President Jacques Chirac (138) President Chirac (4), President Jaques Chirac (3)
Vladimir Putin	Vladimir Putin (146), President Putin (4) President Vladimir Putin (136)
Abdullah Gul	Abdullah Gul (84), Minister Abdullah Gul (57) Prime Minister Abdullah Gul (47), Premier Abdullah Gul (9) Minister Abdullah (74)
Jiang Zemin	Jiang Zemin (94), President Jiang (85), President Jiang Zemin (85), General Secretary Jiang Zemin (2)
John Paul	John Paul (135), John Paul II (57), John Paul II (42)
Silvio Berlusconi	Silvio Berlusconi (113), Prime Minister Silvio Berlusconi (81) Premier Silvio Berlusconi (22), Prime Minister Sivlio Berlusconi (1)
David Beckham	David Beck (94), David Beckham (93), captain David Beckham (17), captain Beckham (5)
Gray Davis	Gray Davis (109), Gov. Gray Davis (73), Governor Gray Davis (26)
Hans Blix	Hans Blix (168), Inspector Hans Blix (17), Dr. Hans Blix (13), U.N. Hans Blix (3)