Pattern Recognition 43 (2010) 1717–1735

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Interesting faces: A graph-based approach for finding people in news

Derya Ozkan [a], Pınar Duygulu [b,*]

[a] University of Southern California, Computer Science Department, Los Angeles, CA, USA
[b] Bilkent University, Department of Computer Engineering, Ankara, Turkey

ARTICLE INFO

Article history:
Received 10 September 2008
Received in revised form
23 September 2009
Accepted 22 October 2009

Keywords:
Face finding
Graph representation
Densest component
Interest points
News photos and videos

ABSTRACT

In this study, we propose a method for finding people in large news photograph and video collections. Our method exploits the multi-modal nature of these data sets to recognize people and does not require any supervisory input. It first uses the name of the person to populate an initial set of candidate faces. From this set, which is likely to include the faces of other people, it selects the group of most similar faces corresponding to the queried person in a variety of conditions. Our main contribution is to transform the problem of recognizing the faces of the queried person in a set of candidate faces to the problem of finding the highly connected sub-graph (the densest component) in a graph representing the similarities of faces. We also propose a novel technique for finding the similarities of faces by matching interest points extracted from the faces. The proposed method further allows the classification of new faces without needing to re-build the graph. The experiments are performed on two data sets: thousands of news photographs from Yahoo! news and over 200 news videos from TRECVid2004. The results show that the proposed method provides significant improvements over text-based methods.

1. Introduction

Face recognition is a long-standing and difficult problem (for recent surveys see [16,50,24,17,49,40,37]). 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 data sets. However, more realistic data sets that contain large variations in pose, illumination and facial expression are still a challenge for face recognition studies.

News photographs and videos, as important sources for stories related to people, constitute one of the most challenging data sets for face recognition. However, the faces in news photographs and/or videos (see Fig. 1) are difficult to recognize using traditional methods. Faces in the news are captured in real-life conditions rather than in restricted and controlled environments, and therefore represent a large variety of poses, illuminations and expressions. These photos/videos are taken both indoors and outdoors. The large variety of environmental conditions, occlusions, clothing and ages make the data set even more difficult to recognize. Scaling is also an important issue for such large volumes of data. Unlike handmade data sets, the number of face categories in the news is unlimited, making classifier-based systems unfeasible.

On the other hand, textual information in news is an important cue for finding photographs or videos of a specific person. In general, a person visually appears when his or her name is mentioned in the news. Therefore, the common approach to finding a person is to search his or her name in the associated caption of news photographs or in the associated speech transcript of video shots.

However, such an approach is likely to yield incorrect results since the photos/shots associated with the name may not include that person or any people at all, or may also contain other people (see Fig. 2(a)). Within the limitations of the selected method, detecting faces and eliminating the photos/shots which do not include any face can handle the problem of having no people in the resulting set. A more difficult problem arises when multiple faces are associated with multiple names, since it is not known which face goes with which name.

In news videos, the problem becomes more challenging due to time shift, which usually occurs between the appearance of the name and the visual appearance of the person. For example, a person's name is mentioned while the anchor is introducing the related story, but the person actually appears later in the video (Fig. 2(b)). Therefore, obtaining the faces from the shot temporally aligned with the speech transcript that includes the name usually yields incorrect results, for the most part returning the faces of the anchor or reporter. As a solution, faces in a sequence of shots in

* Corresponding author.
E-mail addresses: derya@usc.edu (D. Ozkan).
duygulu@cs.bilkent.edu.tr.(P. Duygulu)

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.10.015

Fig. 1. Sample faces from news photographs [5] (top), and from news videos [1] (bottom).

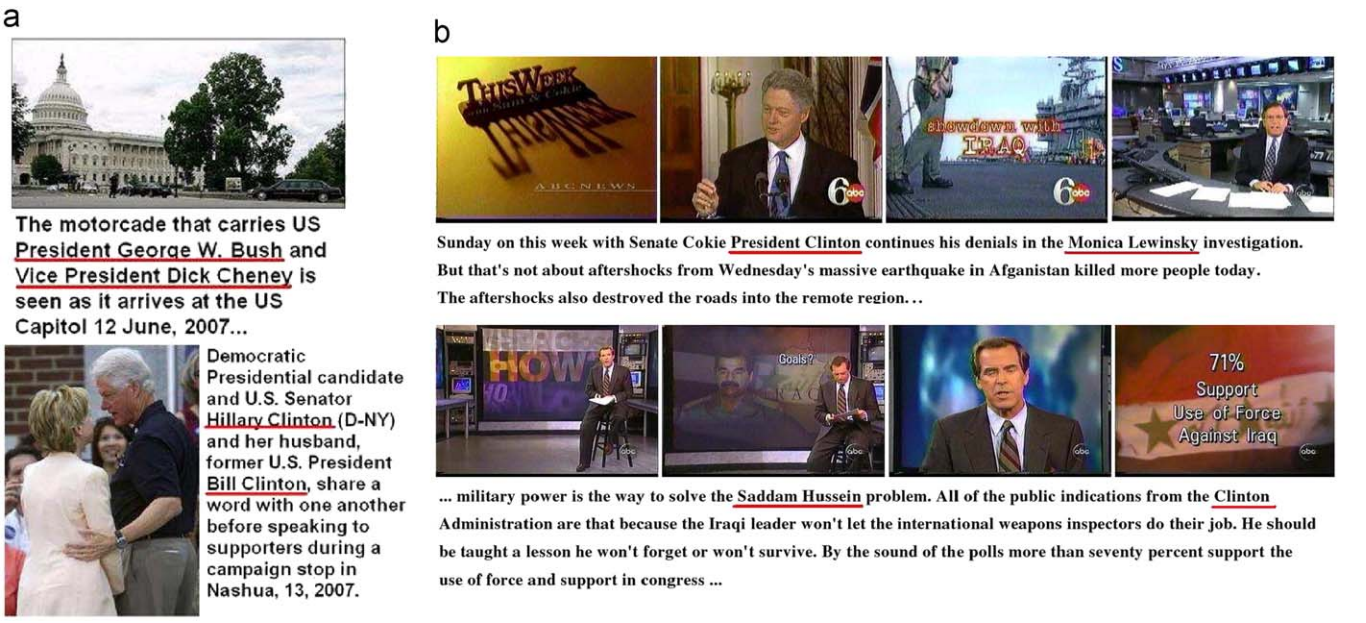


Fig. 2. (a) Sample news photographs and their associated captions. The US president's name may be mentioned while the White House is on the screen (top), or the face of the queried person may appear with other faces, causing an association problem between faces and names (bottom). (b) Sample shots from news videos and the speech transcript texts aligned with those shots. The queried person's name is mentioned when the anchor is on the screen.

the neighborhood of the name can be considered, but this causes large numbers of faces to be associated with a single name, most of which are wrong.

Two important observations should be noted about the results of text-based systems, however. (i) When the faces in the photos/shots associated with a given name are examined, although there may be faces corresponding to other people in the story or some non-face images due to errors of the face detection algorithm used, the faces of the queried person are likely to be the ones most frequently appearing, and (ii) 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 other people. In other words, among a set of faces associated with the name of a person, faces of that person are the ones which are the most similar to each other and forms the largest group of similar faces.

Based on these observations, we approach the problem of finding a person's faces as finding the largest group of most similar faces from a set of faces associated with the name of the person.

To this end, we first find the faces associated with the name of the person. Then, the similarities of these faces are represented in a graph structure, where the vertices correspond to faces and the edges correspond to face similarities. The problem is then turned into finding the densest component of the graph, that is, the largest set of highly connected vertices in the graph, which refers to the largest group of most similar faces corresponding to the faces of the queried person.

The proposed method is not a solution to the general face recognition problem. Rather, on data sets with names and faces together, as in news photos or videos, it is an alternative to supervised and purely vision-based recognition systems—which are likely to fail in such realistic data sets—by making use of both textual and visual cues and using only the available data itself. It is also superior to text-based systems that ignore visual information entirely.

The method does not require a training step and therefore, there is no limitation on the number of people to be recognized and it can be applied to large volumes of data. The only requirement is that the person should be *interesting*, so that a sufficient number of instances are available.

Although the proposed method can be applied to any type of feature extracted from the faces, in this study, a novel method is proposed for matching faces based on the *interest* points extracted from faces, as an alternative to using traditional facial features.

In the following, first we review some of the other studies that use face and name information together. Then, we give a detailed

explanation of the overall graph approach. After describing our data sets, we present detailed experimental results. Next, we compare our proposed method to other approaches. Finally, we conclude with discussions and future directions.

2. Related studies

It has been shown recently that the face recognition problem can be simplified by transforming it into a face-name association problem [36,5,6,13,9].

Satoh and Kanade proposed one of the first studies in [36] and retrieved faces from videos by integrating image understanding and natural language processing. This work is motivated by the fact that humans can identify who is on the screen even if they do not know the person in advance. After finding scene changes by looking for density changes and detecting faces by a neural-network-based method, they first assign similarities among faces by the eigenfaces technique and then extract names by a dictionary-based technique. Finally, they associate names and faces by using co-occurrences of names and faces.

In [47], Yang et al. show that the combination of text and face information allows better retrieval performances in news videos. To reduce the number of faces to be recognized for a queried person, they model the timing between names and appearances of people on the screen. Then, they apply a face recognition method only to the shots selected by the model.

In [12], Ikizler and Duygulu integrate names and faces for improving the retrieval performance of person queries in news videos. Searching the speech transcript text, they first select the key-frames that are aligned with the query name. The faces detected in those key-frames are then clustered and each cluster is represented by one face, to be provided to the users for fast retrieval.

Berg et al. [6] propose a method for associating faces in news photographs with a set of names extracted from the captions. They first reduce the dimensionality using kPCA and LDA processes. Then, they represent each image with both a lower-dimensional image vector and a set of associated names extracted from the caption. Finally, they cluster these images and assign a label for each image. An extended version of this study is presented in [5], which analyzes the language more carefully and uses natural-language to determine who is pictured from text alone.

In [31], text-based results have been improved by searching for faces of other people who occur frequently with the queried person. This process is referred to as query expansion, since the query set is further expanded by looking for the ''friends'' of the queried person who should not be included in the returned results. The authors first apply a generative mixture model to filter the text-based results, then a linear discriminant method for further filtering.

Recently, interest points have been used to represent faces, and Lowe's SIFT descriptor [29] is adapted for this purpose. In [7] application of SIFT approach to face authentication is investigated. Three different matching schemes are proposed in the paper: (i) computing the distances between all pairs of key-points in two images and assigning the minimum distance as the matching score, (ii) using only the features around facial regions such as eyes, and (iii) dividing the face image into grids and matching the features of corresponding grids.

Sivic et al. propose a person-retrieval system in [38] that represents each face image as a collection of overlapping local SIFT descriptors placed at the five facial feature locations (two eyes, mouth, nose and the mid-point between the eyes). The authors 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.

Ballan et al. describe a local feature-based approach for recognizing faces in sports videos [3]. They represent each face by the SIFT features extracted from the five facial feature patches. A similarity metric based on the number of matched interest points is used between two faces. In the paper, they also investigate the relation between the number of matching points and image size, since the resolution might vary in soccer-game videos.

3. Graph-based person finding approach

In this study, we propose a method to finding faces of a person by using the advantages of textual and visual cues in news photographs and videos. Our approach is based on two main assumptions: (i) Although there might be other people or non-person objects/scenes in the photographs or video sequences that include the name of a particular person in the caption or speech transcript, for the most part, that person will be the one who appears visually more than any other individual/object/scene, and (ii) although there could be large variations due to pose, illumination, and scale differences or due to factors such as aging or occlusion, the faces of a particular person tend to be more similar to each other than to faces of other people.

Our approach is based on limiting the faces to be considered to the ones that are associated with the name of the person, and then finding the most similar subset of these faces that is likely to correspond to that person. We use a graph to represent the similarities of faces, and then transform the problem of finding the most similar subset of faces to the problem of finding the densest component in this graph. The densest component corresponds to the set of correct faces of the person, and can be used as a model in classifying new faces. The overall method consists of four steps, as shown in Fig. 3.

The first step consists of a search for a given name in the captions of news photographs or in the speech transcripts of news videos, choosing the photos and video shots that are associated with the query name, and applying a face detection algorithm to the photos and representative key-frames of video shots to find candidate faces.

In the second step, a similarity measure is assigned to each pair of candidate faces and similarities are represented in a graph structure, where vertices are the faces and edges are the similarities between faces. The similarities between faces are represented by using interest points extracted from the faces using Lowe's SIFT operator [29]. We propose a new method for matching the interest points detected in a pair of faces. In this method, initially, all the points on a face are matched with points on another face, based on minimum distance criteria. Then, two constraints, namely a geometric constraint and a unique-match constraint, are applied to eliminate wrong matches and to select the best matching points.

Based on our assumptions, the constructed similarity graph should have two properties: the vertices corresponding to the faces of the person in consideration will be the ones which are more strongly connected to each other than to vertices corresponding to the faces of other people, and the set of such vertices will constitute the largest group in the graph. The set of vertices satisfying these properties corresponds to the densest component of the graph, and found in the third step of the proposed method using a modified version of Charikar's greedy algorithm [8].

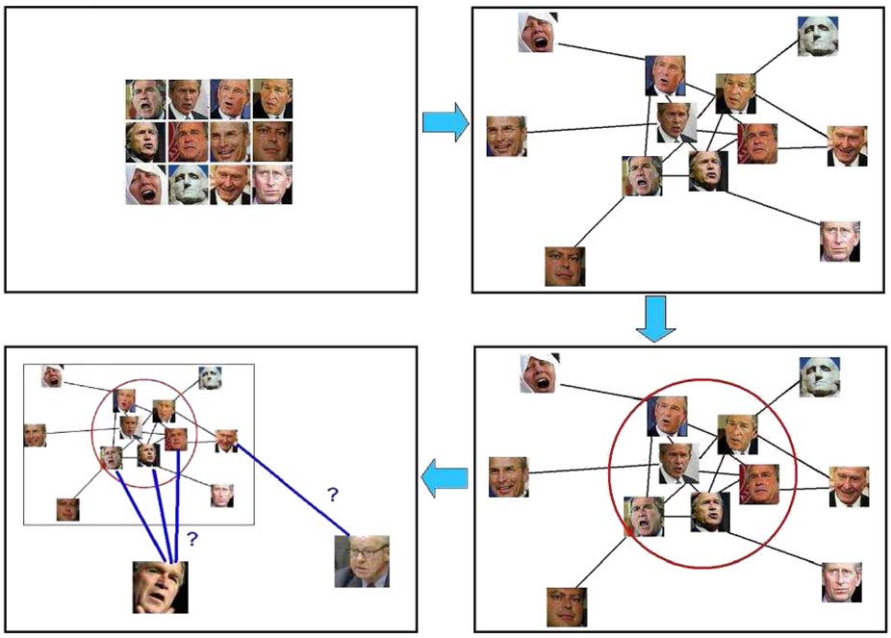


Fig. 3. The four steps of the overall approach are shown for *George W. Bush*. Step 1: applying a face detection algorithm to the photos/video sequences that include the name of a particular person in order to find faces associated with the name. Step 2: defining similarities between these faces to form a similarity graph of faces. Step 3: finding the densest component of this graph corresponding to the set of most-similar faces. Step 4: updating the graph with new data.

Fig. 4. Key-frames from two different videos. The numbers below each image show the distance to the 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 the preceding (left) or succeeding (right) shots.

In the final step, the results of the third step are used as a model to update the graph for inclusion of new data without rebuilding the graph. Two different techniques, namely degree modeling and distance modeling, are proposed for this purpose.

In the following, each step will be described in detail.

3.1. Step 1: finding faces associated with the name

In this step, first we use the name information to find candidate photographs and video shots where it is likely that there is a visual appearance of the person in consideration. To this end, we search for a given name in the captions of the photographs and in the speech transcripts of video shots. However, a person's name can appear in different forms. For example, the names *George W. Bush*, *President Bush*, *US President*, and *President George Bush*, all correspond to the same person. Although automatic named-entity-detection and name-merging methods could be applied, in this study we select and merge the names manually since these aspects are not our focus.

Then, we apply a face detection algorithm to the candidate photographs and representative key-frames of the candidate video shots to remove the ones without any faces and to obtain the set of candidate faces for the given name. Note that there could be more than one detected face in a photograph or in a video key-frame.

In the studies applied to movies and sports videos [38,13,3,2], tracking of faces is performed to obtain multiple faces in a shot. However, in news videos, fast scene changes result in short shot lengths, and therefore tracking is not as advantageous as in other data sets. Thus, to eliminate complexity, we prefer to extract faces only from single key-frames.

In news videos, although it seems likely that a person would visually appear when his or her name is mentioned in the speech transcript text, in most cases, the name is mentioned when the anchor introduces the story, but not when the person appears (see Fig. 4). Therefore, although looking for the shot in which the name is mentioned is a good place to start, it is not sufficient.

Recently, it has been shown that the frequency of a person's visual appearance with respect to the occurrence of his or her name can be assumed to have a Gaussian distribution [47]. We use a similar idea and also consider shots in a close neighborhood of the shot associated with the name as the candidates.

Although this solution reduces the number of missed faces, still one more problem exists: The video key-frames associated with the name mostly include the queried person but also many instances of the anchor. As a solution to this problem, we make use of the proposed method to detect anchors (rather than using a classifier-based anchor detector), and then we remove those faces from the set of faces associated with the name (see Section 3.5 for further details).

3.2. Step 2: constructing the face-similarity graph

After finding candidate faces associated with the name, the next step is to construct a graph representing the similarities of the faces. This step requires (i) a good descriptor for faces and (ii) a good similarity measure.

It has been shown in the literature that [20], component or part-based methods [23,22,30,27,46] which extract features from facial areas are superior to global methods [42–44,4,48] which extract features from the whole face for recognizing faces when there are pose variations, as in the data sets that we consider. On the other hand, the facial-feature extraction step may result in an extra level of noise and may not work in the presence of occlusion or when the deformations are large [38].

In this study, we follow the idea of part-based approaches, but as an alternative to facial features we represent faces with interest points extracted from the faces. Using interest points, which have been recently shown to be successful in recognizing objects, scenes and faces [29,32,21,38,7], our method exploits from their scale and illumination invariance characteristics and can work even in the presence of occlusion. Note that interest points are not only replacements for facial features, but they can also capture specific characteristics that a particular person has, such as a scar.

We use the Difference of Gaussian method to detect the interest points and represent them using the SIFT descriptors [29]. While many interest points are detected, the original matching criteria proposed in [29], which considers the ratio between the best match and the second-best match, results in only a few matches and cannot capture the similarities of faces taken in different conditions, especially with a large variety in pose.

In this study, we propose a new matching method to overcome the problem. We first use a minimum-distance metric to find all the matching points and then remove the false matches by adding some constraints.

First, a face, A, is represented as a set of points $P^A = \{p_1^A \dots p_N^A\}$, where N is the number of all interest points detected on the face. Then, for a pair of faces, A and B, the interest points on one face are compared with the interest points on the other face and the points having the least Euclidean distance are assumed to be the matching ones. That is, a point on face A, p_i^A, is matched to a point on face B, p_j^B, if the Euclidean distance between the SIFT descriptors of the points, $d(p_i^A, p_j^B)$, is the minimum of all distances $d(p_i^A, p_k^B)$, for $k \in P^B$ and $k \neq j$. The matches for all points on face A are then represented as a set $M^A = \{m_1^A \dots m_N^A\}$ where $m_i^A = p_j^B$ if p_i^A is matched to p_j^B.

However, there will be many false matches among this set as well, since each point on one face will have a matching point on the other face (see Fig. 5(a)). In order to eliminate the false matches, we apply two constraints: the geometric constraint and the unique-match constraint. The geometric constraint expects the matching points to appear around similar positions on the faces when the normalized positions are considered, and eliminates the matches where the points do not fall into close positions on the faces. The 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 following, we give the details of how those constraints are applied. Then, we describe the construction of the similarity graph based on the matching points in face pairs.

3.2.1. Geometric 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.

We define the *geometric distance* between a point p_i^A on face A and its matching point m_i^A on face B as $g_i^A = \sqrt{(g_i^A(x))^2 + (g_i^A(y))^2}$ where

$$g_i^A(x) = \frac{p_i^A(x)}{A_W} - \frac{m_i^A(x)}{B_W},$$

and

$$g_i^A(y) = \frac{p_i^A(y)}{A_H} - \frac{m_i^A(y)}{B_H}.$$

Here, $p_i^A(x)$ and $p_i^A(y)$ refer to the x and y coordinates of the point p_i^A, and $m_i^A(x)$ and $m_i^A(y)$ refer to the x and y coordinates of its matching point m_i^A. A_W (B_W) and A_H (B_H) refer to the width and height of face A (B), respectively.

To eliminate the false matches that are distant from each other, we find the geometric distance between matching points, and then eliminate the ones that do not fit to a model learned for geometric distances. For this purpose, we randomly selected five faces 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 [11,45] to classify a matched pair as true or false according to its geometric distance.

In Fig. 5(b), matches after the application of this geometric constraint are shown. Most of the false matches are eliminated when the points that are far away from each other are removed.

3.2.2. Unique-match constraint

After eliminating the points that do not satisfy the geometric constraints, there can still be some false matches. Usually, the false matches are due either to *multiple assignments*, which exist when two or more points on one face are assigned to a single point on the other face (e.g. $m_i^A = p_k^B$ and $m_j^A = p_k^B$, while $m_k^B = p_i^A$), or to *one-way assignments*, which exist when $m_i^A = p_k^B$ but $m_k^B = p_j^A$. These false matches can be eliminated with the application of another constraint, namely the unique-match constraint, which guarantees that each assignment is a two-way assignment, that is, if $m_i^A = p_k^B$ then $m_k^B = p_i^A$.

An example of the matches after applying the unique-match constraint is given in Fig. 5(c). Fig. 6 shows example matches after applying both constraints for some example face pairs.

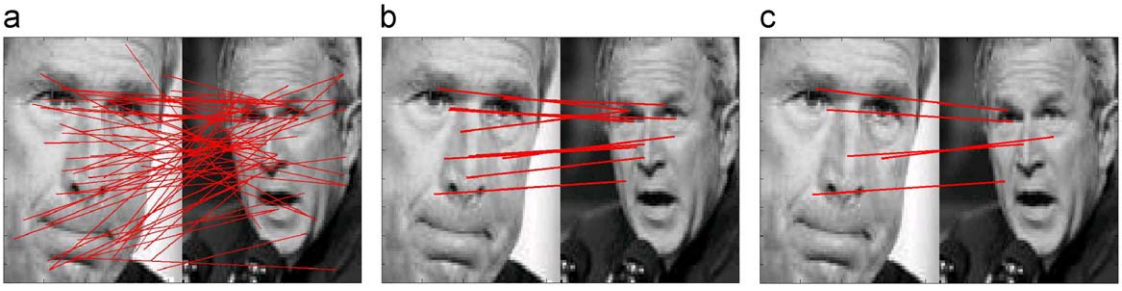


Fig. 5. (a) All the feature points and their matches, based on the minimum-distance criteria. (b) Remaining matches after applying the geometric constraint. (c) Remaining matches after applying the unique-match constraint.

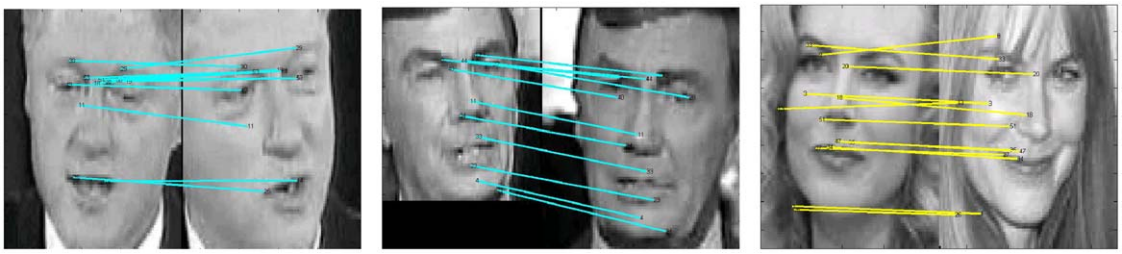


Fig. 6. Sample matching points for faces from news videos after applying all the constraints. Note that even for faces of different sizes, poses or expressions the method successfully finds the corresponding points.

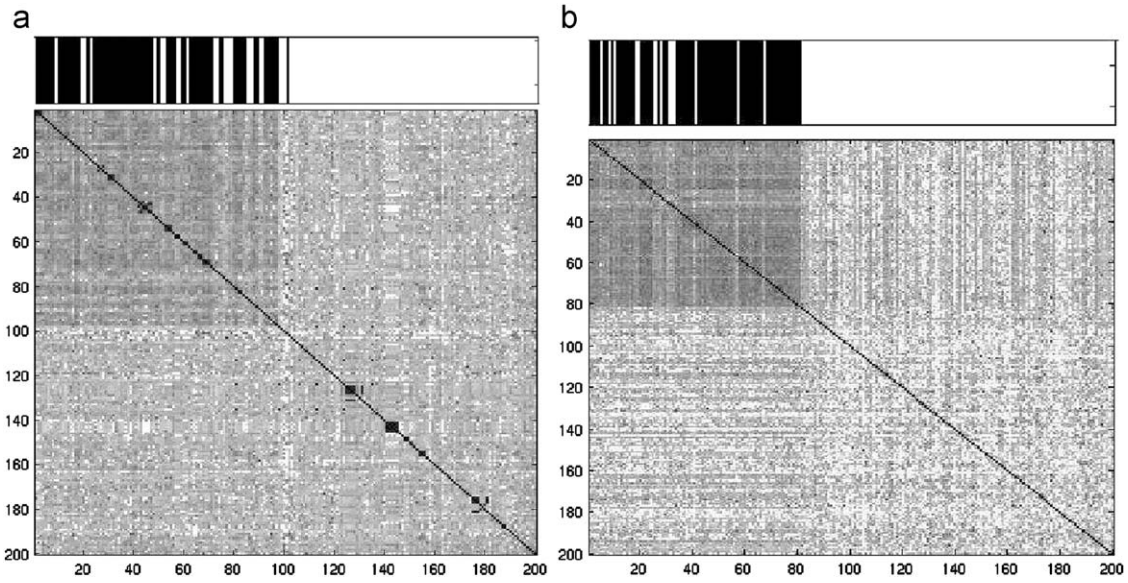


Fig. 7. Samples for constructed similarity matrices. For ease of visualization, the true images are placed on the top left of each matrix. Dark colors correspond to larger similarity values. The bars on the top of the matrices indicate the items found in the densest component of the corresponding graph using the proposed approach. (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.

3.2.3. Similarity-graph construction

After applying the constraints, the distance between two faces A and B is defined through the subset of remaining points, $R^A \subseteq P^A$ and $R^B \subseteq P^B$, as

$$D(A,B) = \frac{\sum_{r_i^A \in R^A} d(r_i^A, m_i^A)}{|R^A|},$$

where $m_i^A \in R^B$ is the match of $r_i^A \in R^A$. Note that $D(A,B) = D(B,A)$.

A similarity graph $G = (V,E)$ for all faces in the search space is then constructed using these distances, where the vertices V correspond to faces, and the edge weights E correspond to the distances between the faces.

We can visualize the graph as a matrix, as in Fig. 7. The matrix is symmetric and the values on the diagonal are all zero. For a clearer visual representation, we rearranged the matrix, to portray together the distances for the faces corresponding to the person we are seeking. Clearly, these faces are more similar to each other than to others. Our goal is to find this subset that will correspond to the densest component in the graph structure.

3.3. Step 3: finding the densest component in the graph

We assume that, in the constructed graph the vertices of a particular person will be close to each other (highly connected) and distant from the other vertices (weakly connected). Hence, the problem can be transformed into finding the densest sub-graph (component) in the entire graph. To find the densest component we adapt the method proposed by Charikar [8].

Let $S \subseteq V$ be the vertices of a sub-graph, and $E(S) = \{i,j \in E : i \in S, j \in S\}$ be the edges of this sub-graph. The density of a sub-graph is defined as

$$f(S) = \frac{|E(S)|}{|S|}.$$

The goal is then to find the sub-graph S that maximizes the density $f(S)$. The method presented in [8] is a greedy algorithm that starts with the entire graph, and removes the vertex with the minimum degree at each step until no vertices remain. That is, initially $S_0 = V$. Then, at each step, v_{min}^i, the vertex with the minimum degree in the sub-graph with S_i is chosen and removed to construct the next sub-graph with $S_{i+1} = S_i - v_{min}^i$. At each step $f(S_i)$ is computed. The algorithm continues until $S_n = \emptyset$. The algorithm then returns the sub-graph S_k with the maximum density $f(S_k) = \max f(S_i)$, $i = 1 \ldots n$ as the densest component of the graph.

The algorithm proposed by Charikar requires a binary graph. In order to apply this algorithm to the weighted graph that we construct, we need to convert our graph into a binary form. Thus, before applying the greedy densest component algorithm, we apply a threshold to edge weights of the original graph

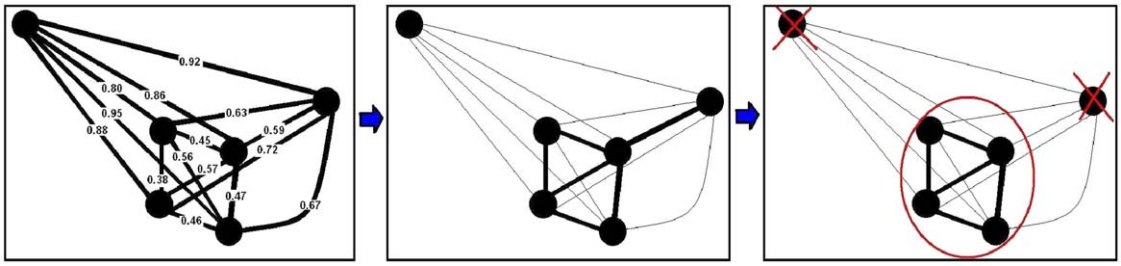


Fig. 8. Example of converting a weighted graph to a binary graph and then finding the densest component. Vertices and their distances are given in the first image. The resulting graph after applying 0.65 as the proximity threshold is given in the second image. Bold edges are the edges that remain after conversion. The final densest component of this graph is circled in the last image.

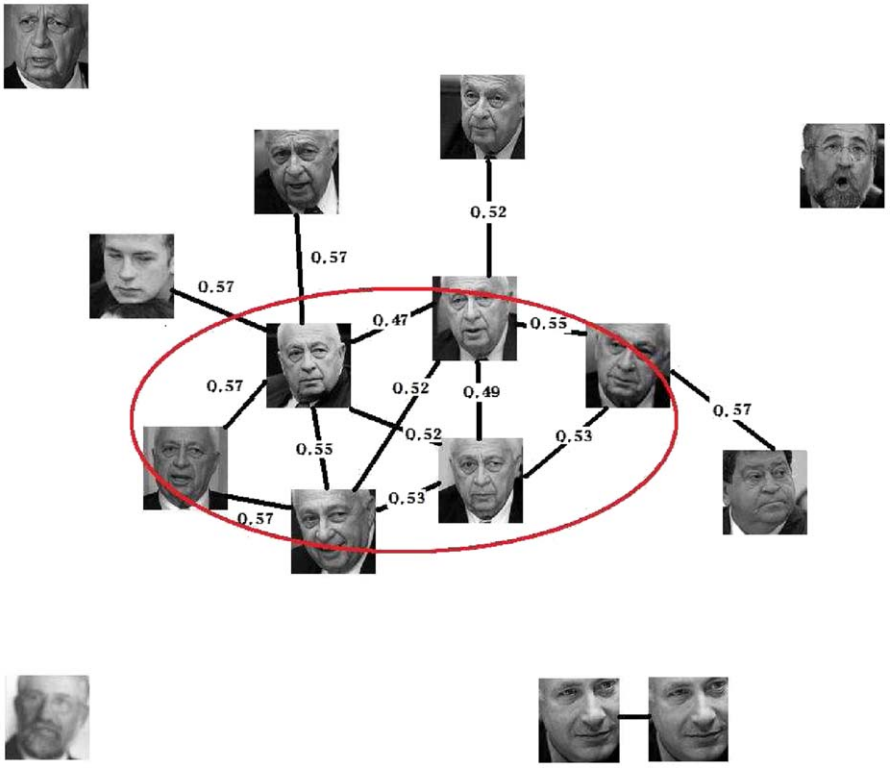


Fig. 9. The links remained after thresholding and the densest component found for a subset of faces associated with Ariel Sharon.

and convert it into a binary form, where 0 indicates no edge and 1 indicates an edge between two vertices. Fig. 8 shows an example of applying the greedy densest algorithm on a weighted graph.

As a real example, in Fig. 9, on a subset of faces obtained for the query *Ariel Sharon* the links remained after thresholding and the densest component found are shown. When all the faces for this query are considered, the average distance between true faces is 0.63, while the average distance between true faces and non-true faces is 0.70. Note that, although some links between true and non-true samples remain after thresholding, since the method searches for strongly connected subset, non-true samples are eliminated since they are only weakly connected to the subset of true faces. Although, some of the true faces are also eliminated, most of the true faces remain inside the densest component despite of large variations in pose, illumination and expression.

3.4. Step 4: dynamic graph update

The overall scheme explained in the previous sections returns a set of faces labeled as the queried person (densest component) and the rest as others (outliers). 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 (queried person or outlier), which is inconvenient when data growth occurs frequently.

We propose a solution to use the resulting graph of the original set as a model to label new faces dynamically. In the next two subsections, we explain two ways to update the graph for the new data. A representation of both methods is depicted in Fig. 10.

3.4.1. Degree modeling

As explained in Section 3.3, the greedy densest component algorithm works iteratively by removing one vertex from the graph until no vertices are left. The average density of each sub-graph is computed in each iteration and finally, the sub-graph with the largest average density is assigned as the densest component. At each iteration $i = 1 \ldots n$, v^i_{min}, the vertex with the minimum degree is removed from the sub-graph to obtain the next sub-graph. The vertex v^k_{min} removed just before reaching the densest component with S_k can be thought of as the breaking point, and indicates evidence of the maximum number of total

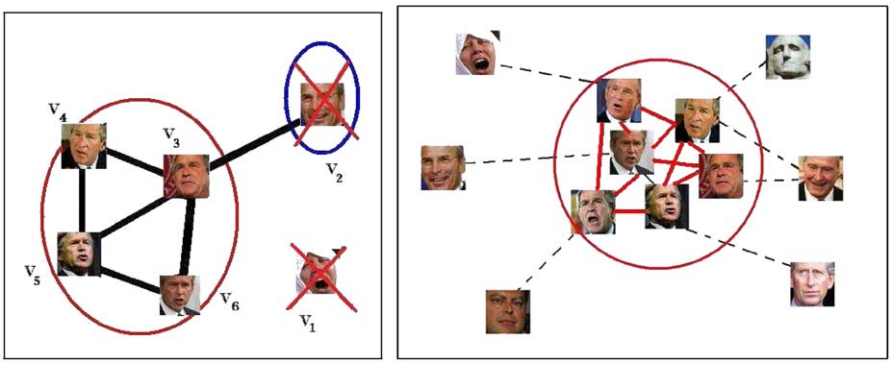


Fig. 10. (Left) An example of how we choose the degree threshold for new faces. V_3, V_4, V_5 and V_6 form the final densest sub-graph. V_2 is $v^k_{\min}$ here, since removing V_2 yields the densest sub-graph while iterating. Hence, the degree threshold is chosen as 1. (Right) An example of how the system learns the distances. Black-dashed edges provide examples of distances to the wrong faces and red edges provide examples of distances to the correct faces. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

connections (edges) from an outlier vertex to all the vertices in the densest component.

We therefore propose to use the degree of the nearest outlier v^k_{min} to all the vertices in the densest component as a threshold for examining a new vertex. When new data is encountered, it is added to the densest sub-graph if its degree is larger than the degree of the last vertex removed, otherwise it is set as an outlier.

3.4.2. Distance modeling

As another method for adding new data items to the densest component we use the weights in the original graph, and model the average distance D_d between the vertices of the densest component, and also the average distance D_o between the vertices of the densest component and the outliers. Formally, for $e(v_i, v_j)$ representing the edge weights between the vertices v_i and v_j, S_k being the subset of the vertices in the densest component, and $G - S_k$ being the subset of the outliers,

$$D_d = avg\, e(v_i, v_j) \quad \text{for } v_i, v_j \in S_k,$$

and

$$D_o = avg\, e(v_i, v_j) \quad \text{for } v_i \in S_k \text{ and } v_j \in G - S_k.$$

For each vertex in the original graph, the average distances for all nodes are computed and then a quadratic Bayes normal classifier [11,45] is used to learn an average distance threshold and to classify new data items based on its average distance to the training examples.

3.5. Anchor detection and removal in news videos

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

In [9], a supervised method for anchor detection is proposed. They integrate color and face information 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 analysis of the audio data. The color information is useful to capture the characteristics of studio settings where the anchor is likely to appear, but when the anchor reports from another environment, this assumption fails. Finally, the method depends on the fact that the faces of the anchor appear in large sizes and around specific positions, but again there may be times when this is not the case.

In this study, we use the graph-based method to find anchors in an unsupervised way. The idea is based on the fact that anchors are usually the most frequently appearing people in broadcast news videos. On different days there may be different anchors reporting, but generally there is a single anchor for each day. Hence, we apply the densest component-based method to each news video separately, to find the person appearing most frequently on that day, which corresponds to the anchor. Then, we apply the proposed method again to the set of all remaining videos where the anchors are excluded to find the faces of the queried person.

4. Experimental results

The method proposed in this study is tested on two different data sets: news photographs on the web [5] and broadcast news videos [1]. In this section, we first briefly describe both data sets used in our experiments, and then evaluate the proposed method on the two data sets separately by discussing several parameters affecting the performance.

Then, in the following sections, we provide comparative experiments to discuss the advantages of the proposed approach over other approaches. First, in Section 5 we focus on the two main contributions of the proposed method, namely selecting features to find similarities of faces, and the graph method as an unsupervised way of finding the most similar faces, and we compare these elements with alternatives. Finally, in Section 6, we compare the proposed method with state-of-the-art studies trying to solve similar problems.

4.1. Data sets and evaluation criteria

4.1.1. News photographs

The data set constructed by Berg et al. originally consisted of approximately 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 [5]. 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 (9609) 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 frequencies 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 manually 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.

Generally, the number of faces in the resulting set is fewer 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 found is 2849, while the total number of all referred names is 7528.

In the experiments, the 23 people with the highest frequency of appearance (more than 200 times) are used. Fig. 11 shows the total number of faces associated with a given name and the number of correctly identified faces for the 23 people used in the experiments.

4.1.2. News videos

The second data set used in the experiments is the broadcast news videos provided by the National Institute of Standards and Technology in USA (NIST) for the TRECVid 2004 video retrieval evaluation competition [1]. It consists of 229 movies of 30 min each from ABC and CNN news. The shot boundaries and key-frames are provided by NIST. Speech transcripts extracted by LIMSI [14] are used to obtain the associated text for each shot.

For the experiments, we choose five 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.

We use a single key-frame from each shot and the face detection algorithm provided by Mikolajcyzk [33] is applied to extract faces from key-frames. Due to high noise levels and low-quality image resolution, the face detector produces many false alarms. Of 10 randomly selected videos, with 2942 key-frames, 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 data set.

For a better understanding of the distribution of the faces around an associated name, we plot the frequency of faces relative to the position of the names for five people, as shown in Fig. 12. As shown in Table 1, when we choose $[-10, 10]$ neighborhood (10 preceeding and 10 following shots), we do not lose many correct faces of the specified person, but the number of faces in this range is very large and includes many false positives. Therefore, we choose a small $[-1, 2]$ neighborhood, and consider only one preceding and two following shots to limit the number of faces without losing many correct faces.

4.1.3. Evaluation criteria

For evaluation purposes, 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 queried person. Among those N, let C be the total number of faces that really belong to that person. Then, the precision value of this result is C/N. If there is a total of M faces in the whole data set that belong to the queried person, then the recall value of the result is C/M. We should also note that as the baseline, we use the images returned by the face detector. Hence, M (ground truth of the queried person) is computed from those detected faces.

After finding the recall and precision values for each queried person individually, we finally compute the weighted results for the average recall and precision values. What we mean by "weighted" is that the recall and/or precision value of each person is weighted by the total number of faces that appear in his or her limited search space. Let $r(i)$ be the recall value for the ith person, and $t(i)$ be the total number of faces in its limited search space. Then, the weighted average recall is

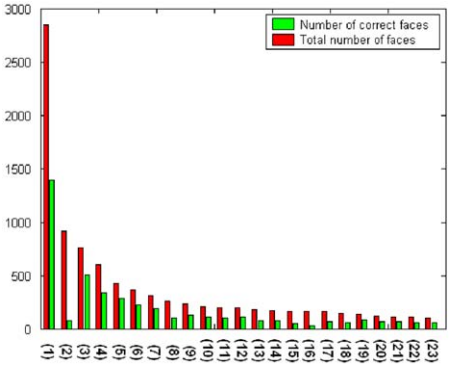
$$weightedAvgRecall = \frac{\sum t(i)r(i)}{\sum t(i)}.$$

The weighted average precision is defined similarly. All the average results given in the rest of the paper refer to the weighted averages.

4.2. Evaluation of the proposed method

4.2.1. Experiments on news photos

The experiments on news photos are first performed on the 23 people appearing with the highest frequencies. Recall and precision values for each person are given in Fig. 13(a). The average precision value is 48%, using the text-based baseline method, which assumes that all faces appearing around the name are correct. With the proposed method, we achieve 68% recall and 71% precision values on average. The method can achieve up to 84% recall as with *Gray Davis* and 100% precision as with *John Ashcroft, Hugo Chavez, Jiang Zemin* and *Abdullah Gul*. We had initially assumed that after associating names, faces of the queried person would appear more than any other person in the search space. Low values are obtained when this is not the case, as with *Saddam Hussein*. There are a total of 913 faces associated with the name *Saddam Hussein*, but only 74 of them are *Saddam Hussein* faces, while 179 of them are *George Bush* faces. Some sample faces retrieved and not retrieved for some people from the test set are shown in Fig. 13(b).



1-George Bush 2-Saddam Hussein
3-Colin Powell 4-Tony Blair
5-Donald Rumsfeld 6-Gerhard Schroeder
7-Ariel Sharon 8-John Ashcroft
9-Hugo Chavez 10-Jean Chretien
11-Hans Blix 12-Junichiro Koizumi
13-Vladimir Putin 14-Jacques Chirac
15-General Kofi 16-John Paul
17-Lula da 18-Gray Davis
19-Silvio Berlusconi 20-David Beckham
21-Jiang Zemin 22-Roh Moo-hyun
23-Abdullah Gul

Fig. 11. (a) Names of 23 people used in the experiments. (b) The total number of faces associated with a name (red bars) and the number of correct faces (green bars). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

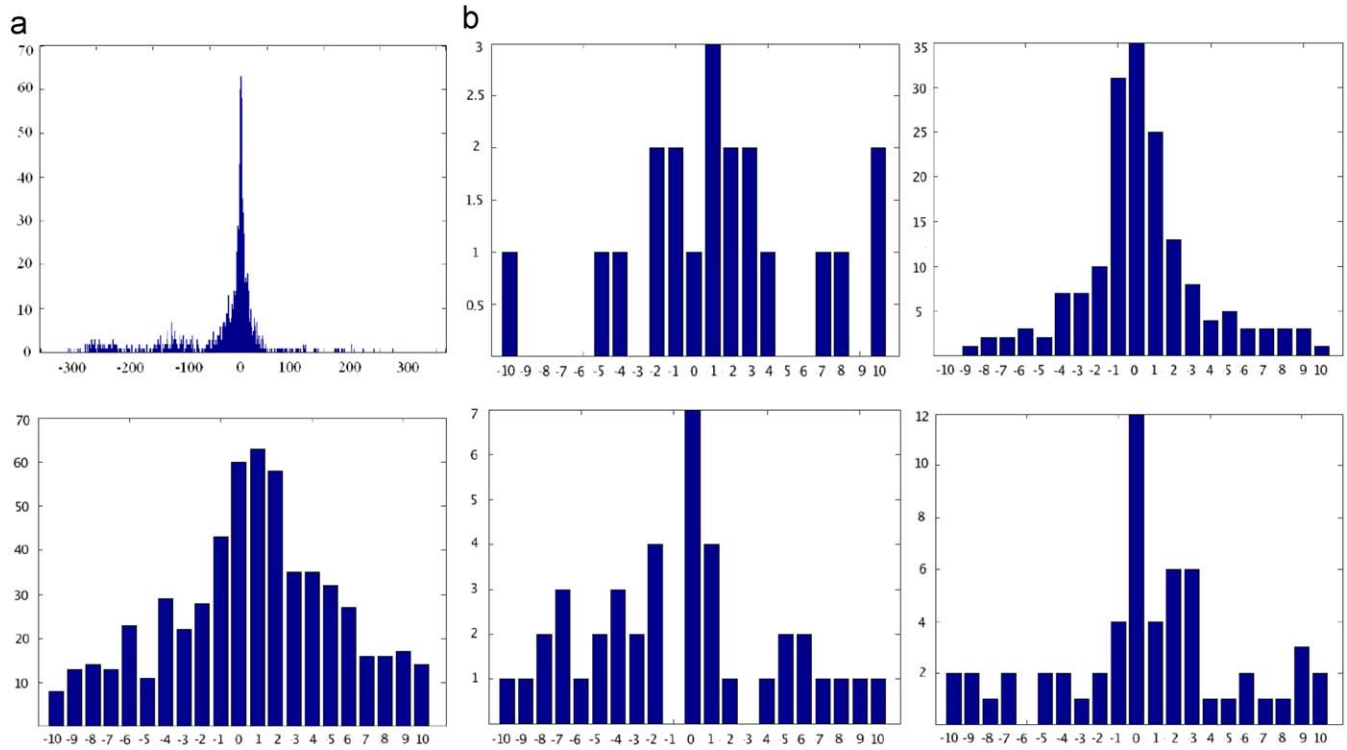


Fig. 12. (a) The figure shows the frequency of Bill Clinton's visual appearance with respect to the distance from the shot in which his name is mentioned (top) when the whole data set is considered, (bottom) when the faces appearing around the name within the preceding and the following 10 shots are considered. Over the whole data set Clinton's face appears 240 times and 213 of them appear in the selected range. (b) The relative position of the faces to the name in a [-10, 10] neighborhood for Benjamin Netanyahu, Sam Donaldson, Saddam Hussein, and Boris Yeltsin, respectively.

Table 1
Number of faces corresponding to the queried name over total number of faces in the ranges [-10, 10] and [-1, 2].

Range	Clinton	Netanyahu	Donaldson	Hussein	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

4.2.2. Experiments on news videos

For the experiments on the news video data set, first we detect the anchors using the densest component based method, and then remove them from the set of faces associated with the queried names. We then process the remaining faces with the method again to find the correct faces.

We run the algorithm on 229 videos in our test set, and obtain average recall and precision values of 90% and 85%, respectively. Faces detected as anchors in six different videos are shown in Fig. 14.

The precision values for the text-only-based method, after removing the anchors, and after applying the proposed graph algorithm are given in Fig. 15(a). Some sample images retrieved for each person are shown in Fig. 15(b).

As can be seen from the results, our method keeps most of the correct faces (especially after anchor removal), and rejects many of the incorrect faces. Hence, the number of images presented to the user is decreased. Also, our improvements in precision values are relatively high. Overall, over the text based baseline method with 9% performance, we achieve 92% recall and 12% precision after removal of the anchors and 67% recall and 15% precision after applying the proposed approach to the remaining faces.

4.2.3. Selection of threshold

The success of our algorithm varies with the threshold chosen to convert the weighted dissimilarity graph to a binary one. To show the effect of threshold, average recall and precision values are plotted as in Fig. 16(a) for news photos and in Fig. 16(b) for news videos, for varying thresholds between 0.55 and 0.65. Based on these values, the threshold 0.575 is determined for news photos and 0.6 is determined for news videos to represent the recall and precision values for each person. The experimental results presented above are obtained using these values.

4.2.4. Results of the dynamic graph update

The above results are presented for the greedy densest component based graph method without supervision. We use the entire set of faces associated with the name as the starting set, and then find the subset of these faces using the proposed method corresponding to the correct faces. In such a setting we assume that the number of faces is constant, and we construct the initial graph using the similarity of all faces associated with the name.

In order to allow the inclusion of new faces into the graph structure, we propose the dynamic graph update method that is described in Section 3.4. To test the effectiveness of the method we use the news photographs data set, and keep the K percentage of the faces associated with a name as a held-out set while constructing the graph with the remaining faces to learn a model. For each K, the algorithm is run 10 times with a different set of random images and the average performance is recorded. The results for two different approaches, namely degree modeling and distance modeling, to finding the similarity of a face from the held-out set to the faces in the densest component are given in Fig. 17. The results show that when there is a sufficient number of

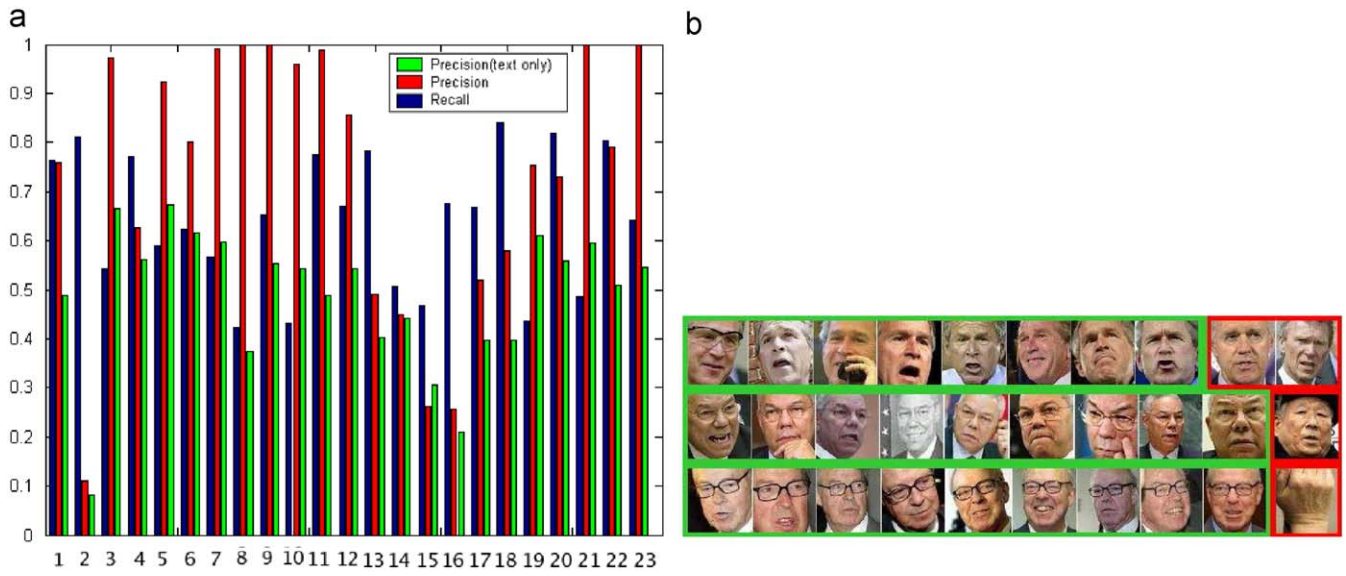


Fig. 13. Results for the news photos data set. (a) Recall and precision values for 23 people for the graph threshold value of 0.575. Blue bars represent recall and red bars represent precision values achieved with the proposed method. Green bars represent precision values for the baseline method, which does not use the visual information and retrieves faces when the name appears in the caption. (b) Sample images retrieved (green) and sample images not retrieved (red) for the query names: *George Bush* (top), *Colin Powell* (middle), *Hans Blix* (bottom). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Fig. 14. Detected anchors for six different videos.

instances for constructing the initial densest component, the inclusion of new data results in almost the same performance, and the results are acceptable even when more than half the data is excluded from the initial set.

4.2.5. *Effectiveness of the method in the case of variations*

In order to test the robustness of our method against variations in the data, such as data size and illumination, additional experiments are conducted.

First, in order to test the robustness of our method in the case of different sample sizes, two experiments on the news photos data set are performed.

In the first experiment, the graph-based method is run on subsets in varying sizes to test the robustness against data size. For this experiment, for each person, among the faces associated with the queried name, K percent of the true faces and K percent of the non-true faces are randomly selected to construct a subset. Ten subsets are constructed in this manner for varying K values, $K = 10, 20, \ldots, 100$. Then, we run the graph based method on these subsets of faces. We use 10-fold cross validation, and hence repeat the experiments 10 times for each K. We average the results of 10 random runs for all people and plot the changes in performance when K changes in Fig. 18. Although there are some slight variations in the overall recall with the changes in the data set size, the precision stays almost constant, in the range of 68% and 72%, after $K = 20$.

As another experiment, we change the number of instances of a face by removing some correct faces or adding some incorrect faces. For four people with approximately 200 instances each and similar numbers of true and false instances, we (i) remove 50 true faces from each search space, and (ii) add 100 false faces. Originally, average recall and precision values were 63% and 95%, respectively. We obtain 59% recall and 89% precision after implementing (i), and 58% recall and 70% precision after implementing (ii). Although the precision is somewhat more negatively affected, results are still acceptable.

Besides the news photo data set, we also perform experiments on the extended Yale Face Database B [15], a benchmark data set for face recognition, to test the robustness against illumination

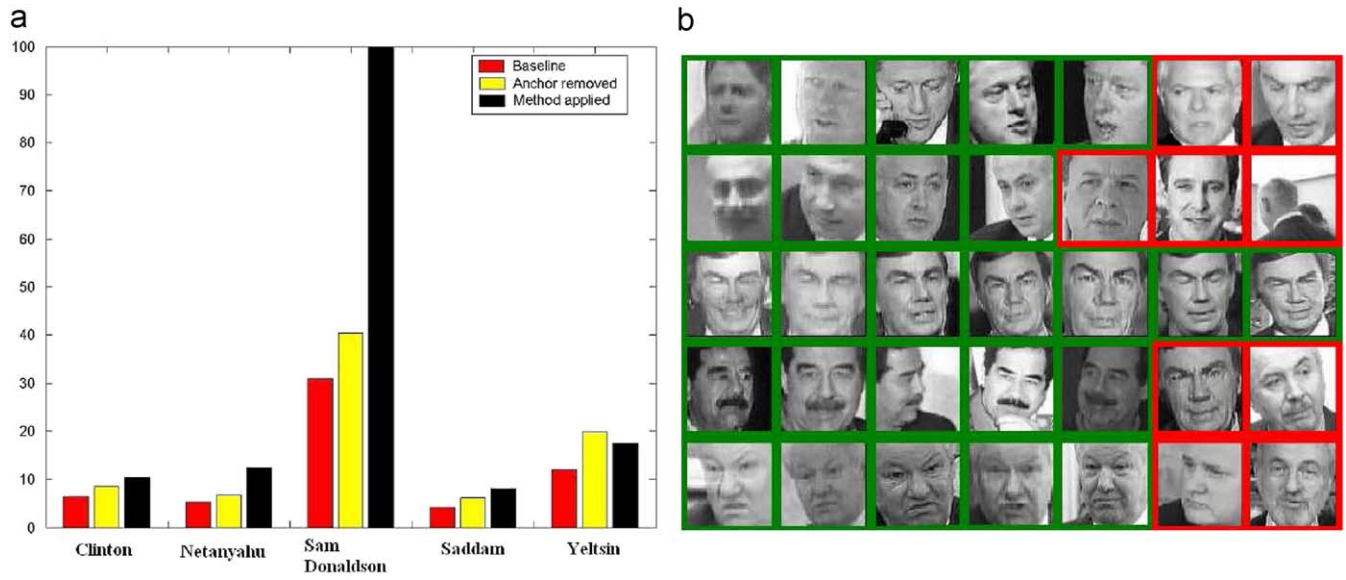


Fig. 15. (a) Precision values achieved for five people used in our tests on news videos. (b) Sample images retrieved.

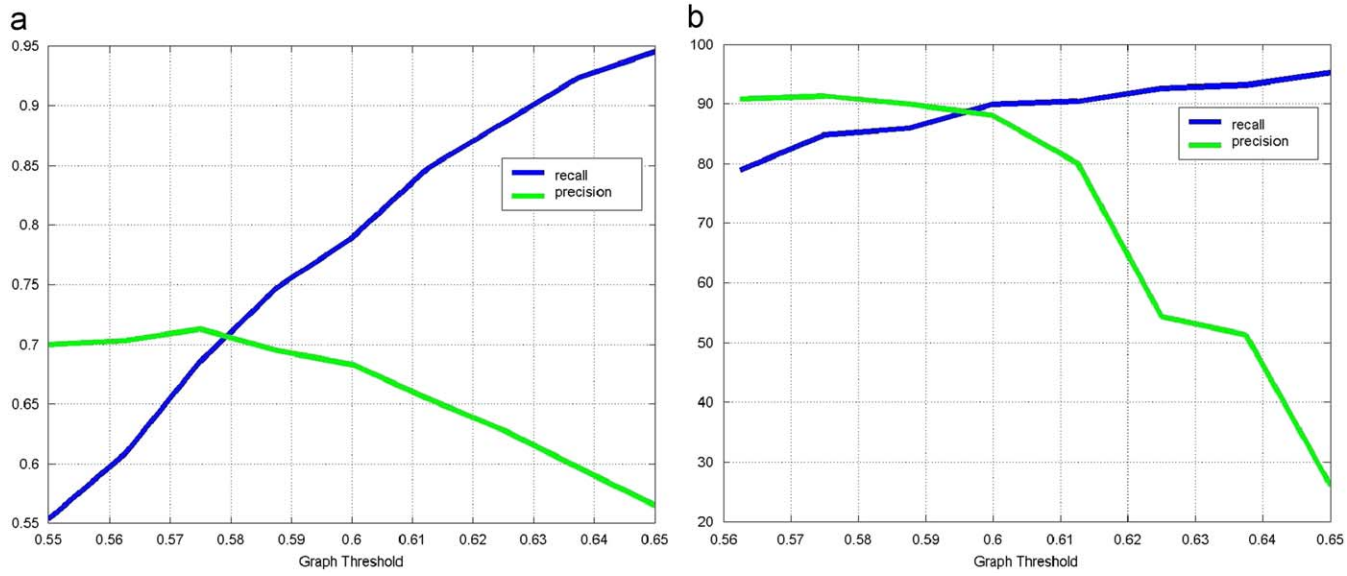


Fig. 16. Average recall and precision values as a function of the graph threshold for (a) news photographs and (b) news videos.

changes. We used a subset of those faces that are manually aligned, cropped, and then re-sized to 168 × 192 [28]. These data contain faces of 38 individuals, all in frontal pose and neutral face expression, but with different lighting conditions. There are two different parameters that determine the lighting: direction of light source with respect to the camera axis (1) in degrees azimuth and (2) in degrees elevation. Illumination is more sensitive to elevation than to azimuth. The higher the degrees, the more the face is subject to illumination.

For this experiment, initially, for each individual, as the true faces, we randomly select faces with degrees azimuth ≤ 20 and degrees elevation ≤ 10. Keeping the degrees same, among the rest of the faces in the data set corresponding to other people, we randomly select the same number of faces for the non-true faces. Note that, in such a subset the baseline precision is 50% (half true, half non-true faces). We run our graph based method on this subset. Then, we increase the variation in lighting with the inclusion of new faces with worse lighting conditions. For this purpose at each step we increase degrees azimuth up to 20 and degrees elevation up to 10 and re-run the graph-based method on these larger subsets. Note that, at each step we add same number of true and non-true faces keeping the baseline precision same.

The recall and precision values for each lighting condition are given in Fig. 19. Although values decrease as the illumination increases as expected, especially when precision is considered it is only a slight difference, and over the baseline precision, the improvement is considerable even when the illumination variations are high.

5. Comparison with alternative approaches

5.1. Describing the similarities of faces

To recall the process of the proposed method, first the points having the minimum distance according to their SIFT descriptors

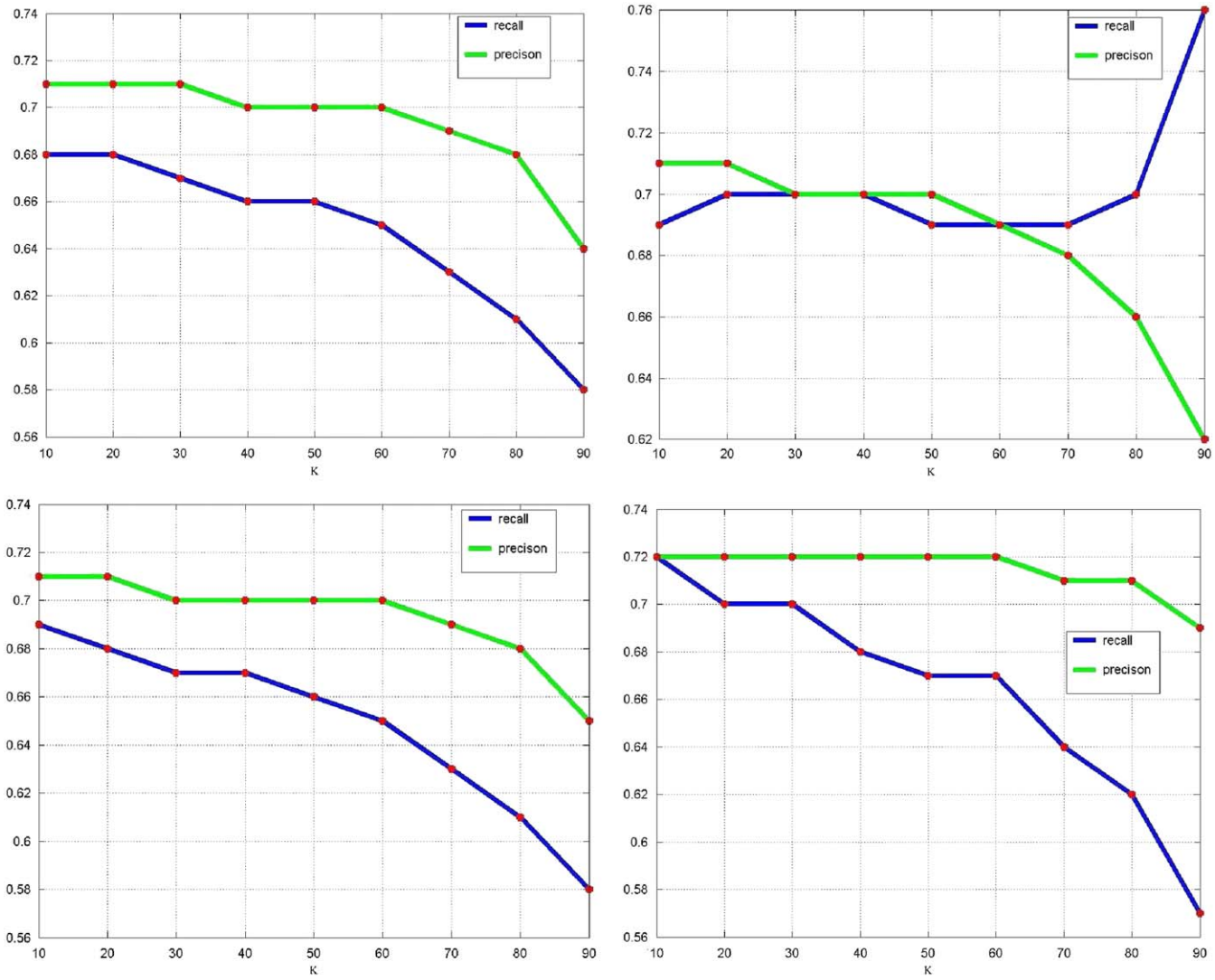


Fig. 17. Average recall and precision values of the model and the held-out sets for the dynamic graph update with degree modeling (top) and distance modeling (bottom). *K* percent of the images are used for held-out.

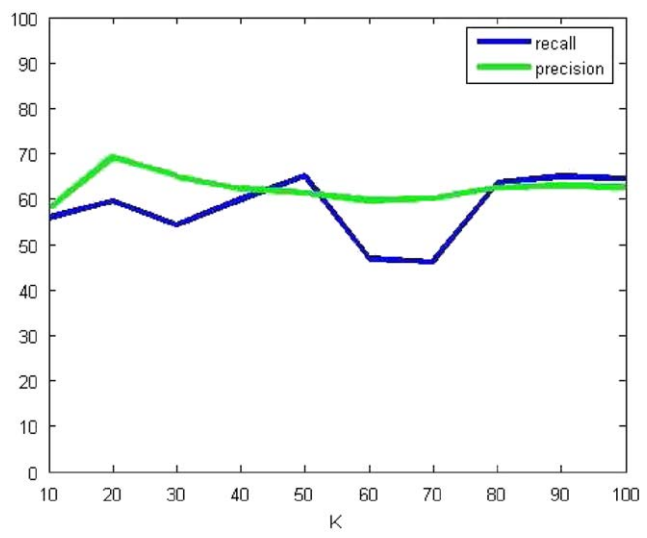


Fig. 18. Performance of our method as the size of data changes. *K* refers to the percentage of the data selected for testing.

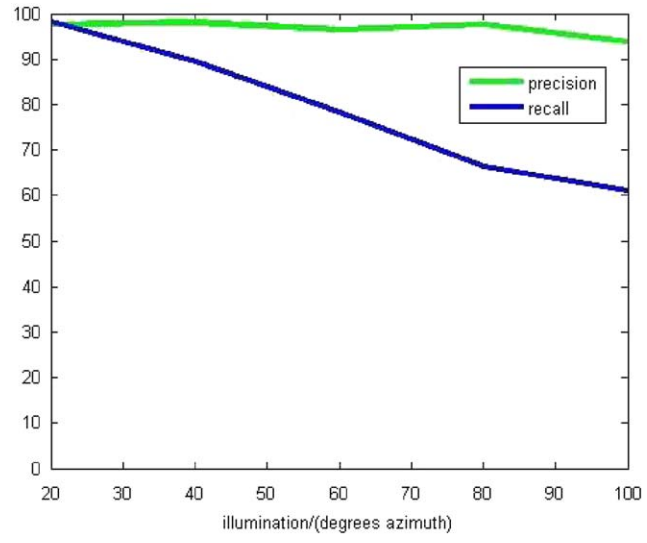


Fig. 19. The recall and precision values as illumination variation increases. At each step, degrees azimuth is changed up to 20 and degrees elevation is changed up to 10.

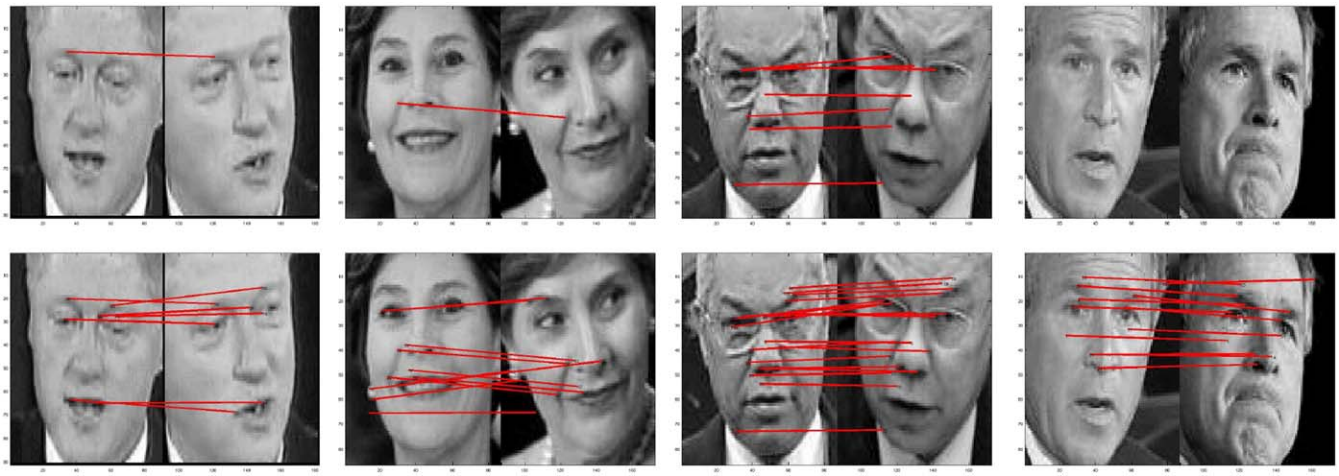


Fig. 20. Examples of matching points. The first row indicates matches found using the original matching metric of Lowe [29]. The second row indicates matches found by applying the proposed method.

are defined as matching points. These points are further eliminated using two constraints, namely the geometric constraint and the uniqueness constraint.

In order to test the effectiveness of the proposed matching method, on a set of faces we manually select the points which should match each other from the set of points initially found using the minimum-distance criteria. We find that after the elimination process, 73% of all possible true matches are kept and only 27% of true matches are lost. Among these assignments, we achieve a correct matching rate of 72%.

As stated earlier, we do not use the original matching metric of Lowe [29], 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 faces are shown in Fig. 20. As the results indicate, the original metric misses many possible matches, while we keep those matches with the proposed scheme.

To see how well the original metric performs, we construct a similarity graph using the matches of the original metric. Then, we apply the densest component graph algorithm to this similarity graph. The recall and precision values for the same threshold in the original tests (0.575) are recorded as 91% and 59%, respectively, which are 68% and 71% with the proposed approach. Although the average recall value seems to be relatively high, the average precision is lower than the one obtained using the proposed approach.

As noted earlier, many face recognition methods use facial features, which refer to eyes, mouth, nose, etc. For comparison, as shown in Fig. 21(a), we have manually label five regions (eyes, mouth, nose and region between the eyes) as in [3,2], for five people from the news photographs data set (*Hans Blix, Jacques Chirac, Kofi Annan, Gray Davis*, and *David Beckham*).

Then, we represent each facial feature by the interest points detected inside this region, if there are any, and find the similarity matrices by comparing only those points. Average recall and precision values are given in Fig. 21(b,c) for both the facial feature approach and the proposed approach. For those five people, the proposed method achieves better precision values. Although recall values are high for the facial features approach, the precision values tend to be closer to the baseline (text-only precision).

These experiments support that, the interest points that are found using the proposed approach are more successful at capturing the characteristics specific to people which are not required to be only the facial features.

Note that, this approach is different than the use of SIFT descriptors extracted to represent facial features as in [38,13,18]. As discussed previously, extraction of facial features requires an additional step and may provide noisy results especially in the presence of the occlusion and large distortions that heavily exist in our data sets. Our proposed approach, which extracts informative parts from the face in general, is an alternative to facial feature based approaches.

5.2. Labeling faces

We compare the greedy graph algorithm for finding the densest component with two other approaches: k-nearest neighbor (k-NN) and the one-class classification. All experiments are conducted on the news photographs data set, where we achieve 68% recall and 71% precision values, on average.

5.2.1. k-nearest neighbor(k-NN) approach

In these experiments a method similar to the k-NN classification is used. For each face in the test set, we find the distances from that face to all the faces in the training set, and select the nearest k faces (k-neighbors). If the number of true faces is greater than the number of false ones in this k-neighborhood, then the test face is classified as a true face. The tests are conducted with different numbers of training and test sets. Each time, we select 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.

The average results are given in Fig. 22. The highest recall and precision values are 68% and 41%, respectively, when $k=3$. Hence, the results indicate that the greedy densest component algorithm outperforms the k-NN approach.

5.2.2. One-class classification

Given a set of data items, one-class classification methods aim to find a target class against the outliers [41]. In that sense, one-class classification methods are suitable for our task and can be compared with the greedy densest component algorithm [8] for our purpose, where we seek to find only the vertices belonging to the densest component (the faces of the queried person) and assume all others are outliers.

However, the similarity graph constructed as described in Section 3.2 keeps only the distances among faces. Hence, one-class classification methods cannot be applied to this graph. So, to compare the greedy graph densest component method with any

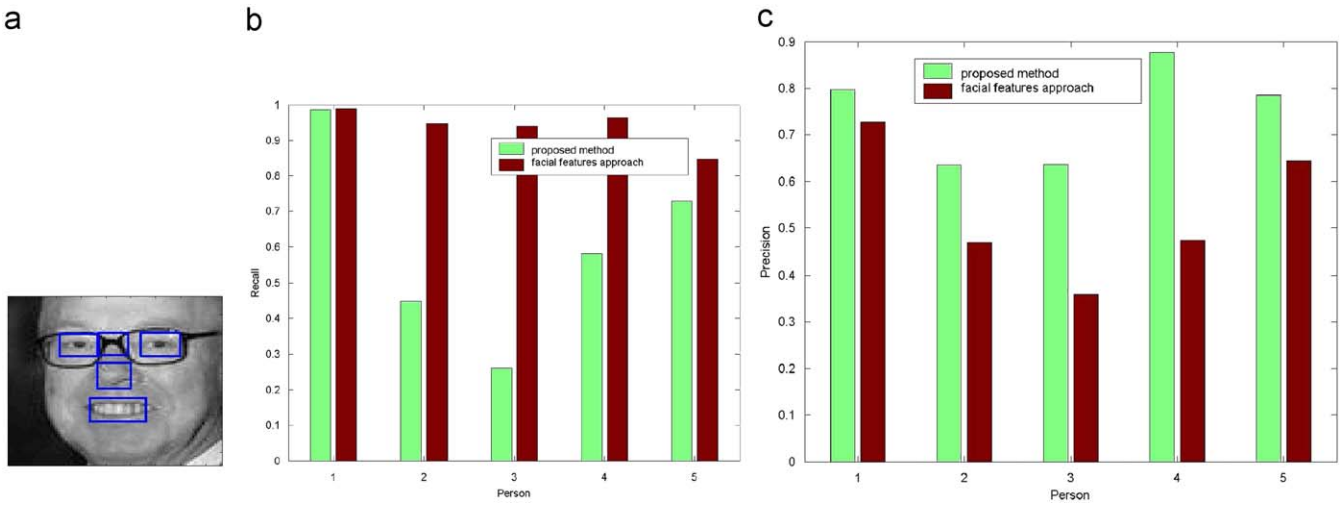


Fig. 21. (a) An example of selected facial regions. (b) Recall and (c) precision values for five people from the news photographs data set using both the facial features and the proposed method.

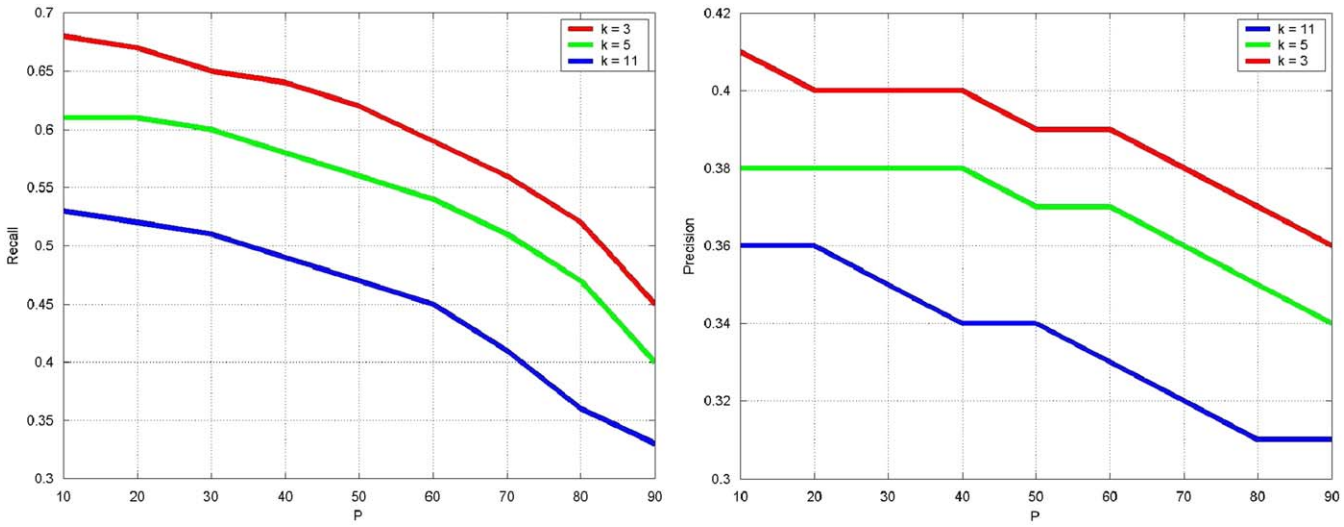


Fig. 22. Recognition rates of the k-NN approach for different P and k values. P is the percentage of images used for testing, and k is the number of neighbors in k-NN.

of those methods, we use the Bag-Of-Features approach, as in [39,26], for graph construction. Then, we apply both the greedy graph method and two different one-class classification methods (nearest neighbor data description method and k-nearest neighbor data description method) giving the best results among other methods.

To construct the graph, we first extract SIFT features from each face and cluster these features using k-means clustering into 50 clusters. Then, a histogram of size 50 is formed for each face, 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

$$tf{-}idf = \frac{n_{id}}{n_d}\log\frac{N}{n_i},$$

where n_{id} is the number of occurrences of term i in document d, n_d is the total number of terms in document d, N is the total number of documents in the database and n_i is the number of documents in the 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 to these features. To apply the densest component graph algorithm, the similarities among the faces are found by the normalized scalar product of tf-idf vectors using the following equation:

$$D_{BOF}(A,B) = \frac{tf{-}idf(A)tf{-}idf(B)}{norm(A)norm(B)},$$

where $tf{-}idf(A)$ is the tf-idf vector of face A and $norm(A)$ is the norm of the tf-idf vector of face A.

The precision-recall curve of the densest component graph algorithm using tf-idf vectors for varying graph thresholds is given in Fig. 23(a).

For one-class classification, K percent of the data is used as held-out and the rest is used as training for K values ranging from 10 to 50, and each step is repeated 10 times. Results are shown in Fig. 23(b,c). For the first one-class classification method (nearest neighbor), for the test set, average recall is approximately 90% while average precision is approximately 50%. Similarly for the second method (k-nearest neighbor), average recall is approximately 86% and average precision is approximately 53%. For recall values of approximately 86% and 90%, we achieve similar precision values as with the one-class classification methods. These results indicate that the one-class classification approach is

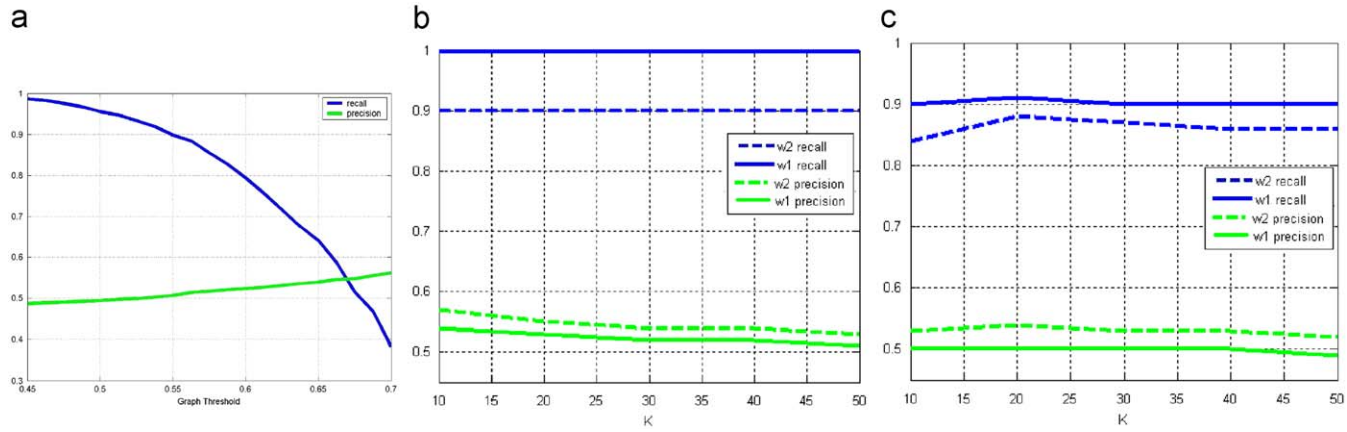


Fig. 23. (a) Recall and precision values as functions of the graph threshold for 23 people in the test set when the greedy densest component algorithm is applied to the features obtained with the Bag-Of-Features approach. (b,c) Recall-precision rates of two one-class classification methods: w1 (nearest neighbor data description method) and w2 (k-nearest neighbor data description method) applied to tf-idf's). (b) Results for training data. (c) Results for test data.

not superior to the greedy densest component algorithm used in our experiments for finding the most similar set of faces.

6. Comparison with related studies

For comparisons with other approaches, we first use the eigenface method as a baseline [42]. The experiments are conducted on the ground truth faces of the top 23 people used in the experiments. For each person, K percent of the faces are kept as held-out and the rest is used for training. 10-fold cross validation is used in each step. Recognition rate is calculated by dividing the correctly labeled faces by the total number of faces. Average recognition rates for training and test sets for different K values are given in Fig. 24. The results show relatively low rates for the test set and reaches only 52% for $K = 10$.

Now, we compare the proposed method with two recent studies. The first study, by Berg et al. [6], associates faces with a set of names. The method is tested on the same data set that we use in the news photographs experiments.

In Fig. 25, recall and precision values of the same 23 people used in our experiments are plotted. In [6] usually better recall values are achieved although we are competitive in most cases. When the precision values are considered, our study performs better in half of the instances and especially when the baseline precision (that is, having a sufficient number of correct faces around the name) is relatively high. In other words, we perform better when our initial assumption holds, that is, when a person appears the most and forms that largest similar group of faces in its limited search space.

The second study used for comparison is a method for re-ranking the results of a search engine by exploiting both keyword and content-based retrieval [19]. This study is chosen because of the relevancy in selecting the required instances with a clustering method.

The method 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 [10] and finds the cluster of blobs corresponding to the largest number of parent images. This cluster is called the significant cluster, and its mean is then used for re-ranking all the search results based on the distance of each blob to the significant cluster.

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

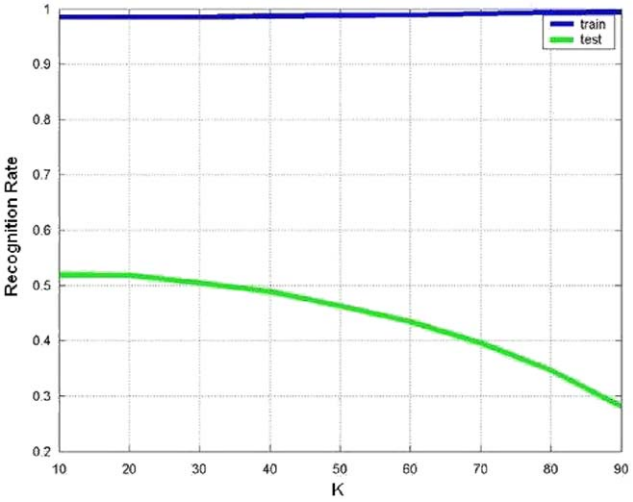


Fig. 24. Recognition rates of the eigenface method for training and test sets as a function of K (the percentage of the held-out set).

In order to be able to compare the two studies, we first need to convert our representation—which is in the form of similarities between vertices—into a vector form. For this purpose, we apply a multi-dimensional scaling method [25] on the weighted similarity matrix of the 23 people in the news photographs data set and find the x–y coordinates of each item (faces) on a plane. Using their x–y coordinates, we cluster these images also by the mean-shift clustering method used in [10] and find the significant cluster corresponding to the faces of the queried 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 plot the recall-precision values as a function of the window size in Fig. 26(a). Although the recall value can increase up to 100%, the maximum precision value we obtained is 51%, which does not improve on the text-only results (48%).

In a second experiment, we apply the significant clusters by applying the mean-shift clustering method on the features obtained by the Bag-Of-Features approach described in the

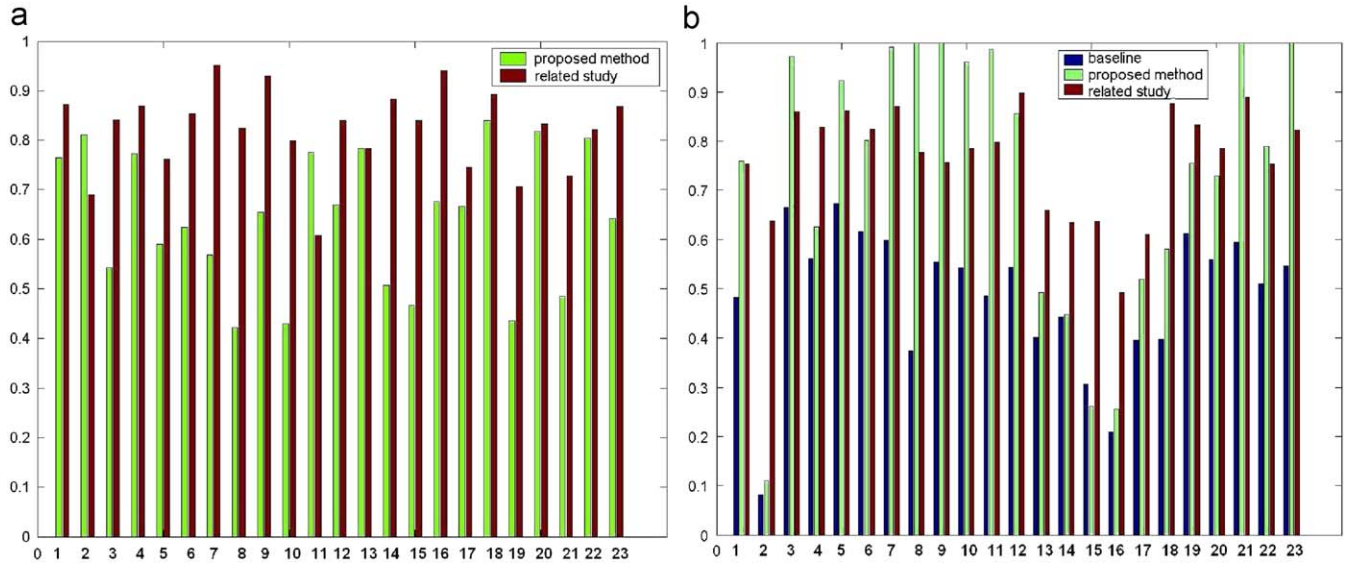


Fig. 25. Comparison with the study of Berg et al. [6] using (a) recall and (b) precision values for the 23 people from the news photographs data set.

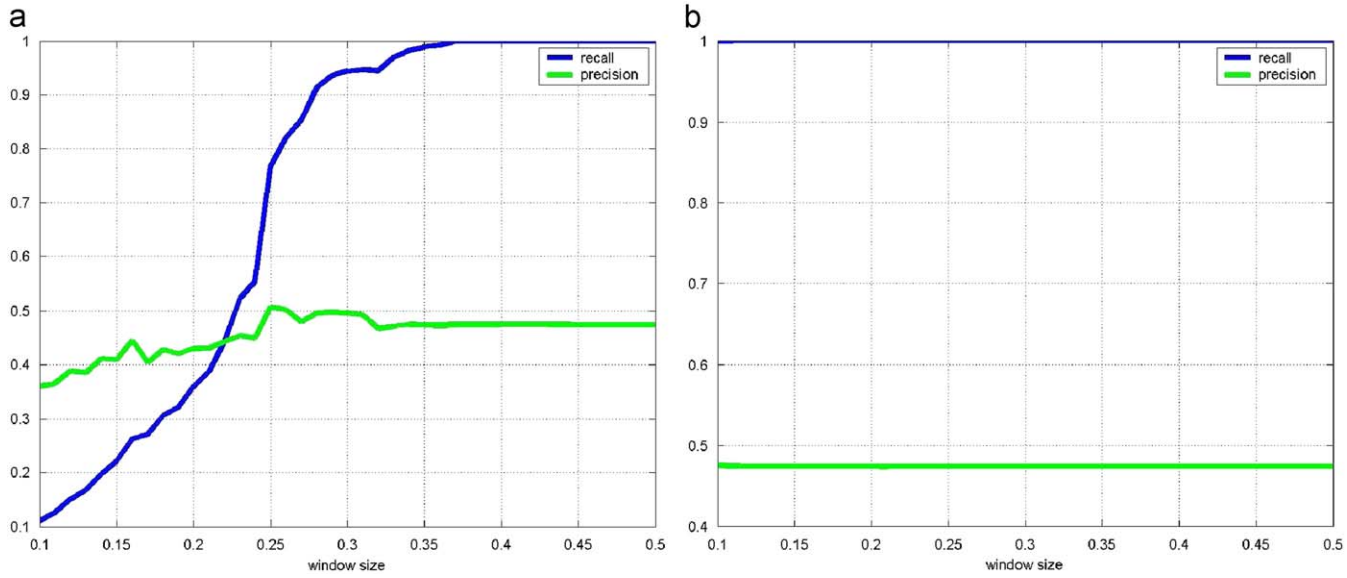


Fig. 26. Recall and precision values for 23 people in the news photographs data set as a function of the window size used in mean-shift clustering, when the significant cluster approach is applied to (a) multi-dimensional scaling coordinate features, and (b) Bag-Of-Features representation.

previous section. Similarly, although the recall value obtained is around 100%, the precision value remains around baseline (48%); hence there is no improvement on the text-only results (see Fig. 26(b)).

In [18], Guillaumin et al. add another constraint over our base method reported in [35,34], based on an assumption that limits each image associated with the name to contain only a single instance of the queried person. Although slight improvements over our base method are obtained on the news photo data set, this assumption has a drawback. Recall that in news videos, the face does not usually appear in direct relation to the transcript, but in the neighborhood of its mention, and it is likely that there will be more than one face of the same person in that neighborhood. Limiting the number to one, is likely to result in losing important data for constructing the graph and therefore result in a decrease in performance.

7. Summary and discussions

In this paper, we propose a method for finding people in large news photograph and video collections where it is difficult to apply traditional face recognition methods because of variations in pose, illumination, and expression, and also due to occlusion, and differences in time and clothing.

The proposed method first limits the search space to only the detected faces in the photos or video shots associated with the name of the queried person in the caption or speech transcript text. Then, it applies a graph-based method to incorporate the visual similarities of faces to improve text-based results.

The proposed approach does not require any form of manual labeling and makes use of only the large volume of the data itself. Therefore, there is no limit to the number of faces that can be recognized.

Our major contributions are threefold:

- We transform the problem of finding the most similar faces in a set of faces linked to a name to the problem of finding the densest component in a similarity graph. For this purpose, we exploit Charikar's greedy densest component algorithm [8] with a modification to be applicable to a weighted graph.
- We represent the similarities of faces using the interest points extracted from the faces as an alternative to using facial features. The limitations of the original matching criteria of [29] are handled by introducing two new constraints to provide correct matches between faces capturing the characteristics specific to people.
- We allow the inclusion of new faces, which is important for the realistic data sets as in the news photos and videos that grow rapidly. The constructed graph is used as a model, with the introduction of two novel methods to classify new faces.

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. Compared to the 48% of an only text-based baseline performance, we achieve 68% recall and 71% precision on average and up to 84% recall and 100% precision for some individuals. The second data set consists of 229 broadcast news videos. On this set we first use the proposed approach as a new method for detecting news anchors without any supervision and achieve a performance of 90% recall and 85% precision. Then, in experiments on recognizing the faces of five people, we achieve 92% recall and 12% precision after removing the anchors and 67% recall and 15% precision after applying the proposed approach to the remaining faces. The text-based baseline method achieves only 9% performance.

We show that, the proposed method with the introduction of two new constraints is better at matching interest points than the original matching criteria. We also show that when compared with two supervised methods, the proposed method is superior to k-NN and comparable to one-class classification methods. We also see that our results are comparable to the results of Berg et al., and better when the baseline precision values are high. This satisfies our initial assumption that the most commonly occurring face around the name would be the queried person.

There are still some possible directions to improve the performance of our method. In this study, the weighted similarity graph is converted into a binary graph to apply the greedy densest component algorithm. This results in a dependency on threshold selection and loss of some information. Keeping our original idea of assigning the name to the largest set of most similar faces, another method could be used for finding that subset. Although we propose a novel method for representing faces using matching interest points, the overall method could be applied to any other form of face descriptor or similarity criteria. Similarly, the constraints could be replaced by some other method that considers the topology of faces. In this study, the face detector of [33] is used, as it is in Berg et al.'s work [5]. A more powerful face detector is highly likely to produce better performance. Also, as in some other studies, face tracking can be applied to videos to obtain a set of faces rather than a single face representing a single shot.

The proposed idea can also be applied to other problems such as object recognition or image region annotation, as we show by applying it to web image re-ranking in [51].

Acknowledgements

This research is partially supported by TÜBİTAK projects Grant nos. 104E065 and 104E077. We would like to thank Rana Nelson for proof reading.

References

[1] Trec video retrieval evaluatio, ⟨http://www-nlpir.nist.gov/projects/trecvid⟩, 2004.
[2] L. Ballan, M. Bertini, A.D. Bimbo, W. Nunziati, Automatic detection and recognition of players in soccer videos, in: International Conference on Visual Information Systems (VISUAL), Shanghai, China, 2007, pp. 105–116.
[3] L. Ballan, M. Bertini, A.D. Bimbo, W. Nunziati, Soccer players identification based on visual local features, in: 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 2007, pp. 258–265.
[4] M. Bartlett, H. Lades, T. Sejnowski, Independent component representations for face recognition, in: SPIE Symposium on Electronic Imaging: Science and Technology; Human Vision and Electronic Imaging III, 1998.
[5] T. Berg, A.C. Berg, J. Edwards, D. Forsyth, Who's in the picture, in: Neural Information Processing Systems, 2004.
[6] T. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.W. Teh, E. Learned-Miller, D.A. Forsyth, Names and faces in the news, in: IEEE Conference on Computer Vision and Pattern Recognition, 2004.
[7] M. Bicego, A. Lagorio, E. Grosso, M. Tistarelli, On the use of sift features for face authentication, in: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, 2006, p. 35.
[8] M. Charikar, Greedy approximation algorithms for finding dense components in a graph, in: International Workshop on Approximation Algorithms for Combinatorial Optimization, London, UK, 2000.
[9] M.Y. Chen, A. Hauptmann, Searching for a specific person in broadcast news video, in: International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, 2004.
[10] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (8) (1995) 790–799.
[11] R.O. Duda, P.E. Hart, D.G. Stork, Pattern classification, Wiley, New York, 2001.
[12] P. Duygulu, A. Hauptmann, What's news, what's not? associating news videos with words, in: The 3rd International Conference on Image and Video Retrieval, Ireland, 2004.
[13] M. Everingham, J. Sivic, A. Zisserman, Hello! my name is... buffy – automatic naming of characters in tv video, in: Proceedings of the British Machine Vision Conference, 2006.
[14] J.L. Gauvain, L. Lamel, G. Adda, The LIMSI Broadcast News Transcription System, Speech Communication 37 (1–2).
[15] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 643–660.
[16] R. Gross, S. Baker, I. Matthews, T. Kanade, Face recognition across pose and illumination, in: S.Z. Li, A.K. Jain (Eds.), Handbook of Face Recognition, Springer, Berlin, 2004.
[17] R. Gross, J. Shi, J. Cohn, Quo vadis face recognition? in: Third Workshop on Empirical Evaluation Methods in Computer Vision, 2001.
[18] M. Guillaumin, J.V.T. Mensink, C. Schmid, Automatic face naming with caption-based supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[19] N.B. Haim, B. Babenko, S. Belongie, Improving web-based image search via content based clustering, in: Semantic Learning Applications in Multimedia, 2006, p. 106.
[20] B. Heisele, P. Ho, J. Wu, T. Poggio, Face recognition: compensated versus global approaches, Computer Vision and Image Understanding 91 (1–2) (2003) 6–21.
[21] S. Helmer, D.G. Lowe, Object recognition with many local features, in: Workshop on Generative Model Based Vision, Washington, DC, 2004.
[22] Y. Kaya, K. Kohayashi, A basic study on human face recognition, in: S. Watanabe (Ed.), Frontiers of Pattern Recognition, 1972, p. 265.
[23] M.D. Kelly, Visual identification of people by computer, Ph.D. Thesis, 1971.
[24] S. Kong, J. Heo, B. Abidi, J. Paik, M. Abidi, Recent advances in visual and infrared face recognition: a review, Computer Vision and Image Understanding 97 (1) (2005) 103–135.
[25] J.B. Kruskal, M. Wish, Multidimensional Scaling, Sage University Paper Series on Quantitative Application in the Social Sciences, Sage Publications, Beverly Hills and London, 1978.
[26] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006.
[27] H. Le, H. Li, Recognizing frontal face images using hidden Markov models with one training image per person, in: International Conference on Pattern Recognition, 2004, pp. 318–321.
[28] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (5) (2005) 684–698.
[29] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2).
[30] B.S. Manjunath, R. Chellappa, C. von der Malsburg, A feature based approach to face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 1992.
[31] T. Mensink, J. Verbeek, Improving people search using query expansions: how friends help to find people, in: European Conference on Computer Vision, 2008.
[32] K. Mikolajczk, C. Schmid, A performance evaluation of local descriptors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2003.

[33] K. Mikolajczyk, Face detector, Ph.D Report, INRIA Rhone-Alpes, 2004.
[34] D. Ozkan, A graph based approach for finding people in news, Master's Thesis, Bilkent University, August 2007.
[35] D. Ozkan, P. Duygulu, Interesting faces in the news, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1477–1482.
[36] S. Satoh, T. Kanade, Name-it: association of face and name in video, in: IEEE Conference on Computer Vision and Pattern Recognition, 1997, p. 368.
[37] P. Sinha, B. Balas, Y. Ostrovsky, R. Russell, Face recognition by humans: 19 results all computer vision researchers should know about, Proceedings of the IEEE 94 (11) (2006) 1948–1962.
[38] J. Sivic, M. Everingham, A. Zisserman, Person spotting: video shot retrieval for face sets, in: International Conference on Image and Video Retrieval, 2005, pp. 226–236.
[39] J. Sivic, A. Zisserman, Video Google: a text retrieval approach to object matching in videos, in: IEEE International Conference on Computer Vision, 2003, p. 1470.
[40] X. Tan, S. Chen, Z.H. Zhou, F. Zhang, Face recognition from a single image per person: a survey, Pattern Recognition 39 (9) (2006) 1725–1745.
[41] D.M.J. Tax, One-class classification, Ph.D. Thesis, Delft University of Technology, June 2001.
[42] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
[43] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: IEEE Conference on Computer Vision and Pattern Recognition, 1991, pp. 586–591.
[44] J. Wang, K.N. Plataniotis, A.N. Venetsanopoulos, Selecting discriminant eigenfaces for face recognition, Pattern Recognition Letters 26 (10) (2005) 1470–1482.
[45] A. Webb, Statistical Pattern Recognition, Second ed., Wiley, New York, 2002.
[46] L. Wiskott, J.-M. Fellous, N. Krüger, C. von der Malsburg, Face recognition by elastic bunch graph matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 775–779.
[47] J. Yang, M.Y. Chen, A. Hauptmann, Finding person x: correlating names with visual appearances, in: International Conference on Image and Video Retrieval, Dublin City University Ireland, 2004, pp. 270–278.
[48] J. Yang, D. Zhang, A.F. Frangi, 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) (2004) 131–137.
[49] W. Zhao, R. Chellappa, Image-based face recognition: issues and methods, in: B. Javidi, M. Dekker (Ed.), Image Recognition and Classification, 2002, pp. 375–402.
[50] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Computing Surveys (2003) 399–458.
[51] H. Zitouni, S.G. Sevil, D. Ozkan, P. Duygulu, Re-ranking of image search results using a graph algorithm, in: 19th International Conference on Pattern Recognition (ICPR2008), Tampa, Florida, USA, 2008.

About the Author—DERYA OZKAN has received her BSc, MSc degrees from Department of Computer Engineering at Bilkent University. During her master thesis she worked on recognition and retrieval of faces under the supervision of Dr. Pinar Duygulu. Between June 2006 and February 2007, she worked as a research intern at Siemens Corporate Research, Princeton, NJ. Derya is now a PhD Student at Computer Science Department, University of Southern California (USC) since August 2007 under the supervision of Prof. Gerard Medioni and she is a member of USC Computer Vision Laboratory. She works on 3-D face modeling and recognition.

About the Author—PINAR DUYGULU has received her BSc, MSc and PhD degrees from Department of Computer Engineering at Middle East Technical University, Ankara, Turkey in 1996, 1998 and 2003, respectively. During her PhD, she was a visiting scholar at University of California at Berkeley under the supervision of Prof. David Forsyth. After being a post-doctoral researcher at Informadia Project at Carnegie Mellon University, she joined to Department of Computer Engineering at Bilkent University, Ankara, Turkey. She is the co-director of RETINA Vision and Learning Group at Bilkent. During summer 2004 she was a senior researcher in CLSP summer workshop at Johns Hopkins University on Joint Visual Text Modeling. She received the best paper in Cognitive Vision award at European Conference on Computer Vision in 2002. Her current research interests include computer vision and multimedia data mining, specifically object, face and action recognition and semantic analysis in large image and video collections.