Person Search Made Easy

Nazlı İkizler and Pınar Duygulu

Department of Computer Engineering,
Bilkent University, Ankara, Turkey
{inazli, duygulu}@cs.bilkent.edu.tr

Abstract. In this study, we present a method to extensively reduce the number of retrieved images and increase the retrieval performance for the person queries on the broadcast news videos. A multi-modal approach which integrates face and text information is proposed. A state-of-the-art face detection algorithm is improved using a skin color based method to eliminate the false alarms. This pruned set is clustered to group the similar faces and representative faces are selected from each cluster to be provided to the user. For six person queries of TRECVID2004, on the average, the retrieval rate is increased from 8% to around 50%, and the number of images that the user has to inspect are reduced from hundreds and thousands to tens.

1 Introduction

News videos, with their high social impact, are a rich source of information, therefore multimedia applications which aim to ease their access are important. Indexing, retrieval and analysis of these news videos constitute a big challenge due to their multi-modal nature. This challenge has recently been acknowledged by NIST and broadcasted news videos are chosen as the data set for the TRECVID (TREC Video Retrieval Evaluation) [1] competition.

Broadcast news mostly consist of stories about people making the queries related to a specific person important. The common way to retrieve the information related to a person is to query his/her name on the speech transcript or closed caption text. Such retrieval methods are based on the assumption that a person is likely to appear when his/her name is mentioned. However, this assumption does not always hold. For instance, as shown in Figure 1, Clinton's face appears when his name is not mentioned in the speech transcript, and whenever the anchorperson or the reporter is speaking, his name is mentioned. As a result, a query based only on text is likely to yield frames showing the anchorperson or the reporter.

In order to retrieve the images of a particular person, visual information has to be incorporated and the face of the person needs to be recognized. However, face recognition is a long standing problem, and most results on face recognition methods are evaluated only on controlled environments and for limited data sets [2]. The noisy and complicated nature of news videos makes the face recognition on videos even more challenging.

Currently, there is no fully automatic system to search for specific people in the large image and/or video archives. In most of the existing systems, human

W.-K. Leow et al. (Eds.): CIVR 2005, LNCS 3568, pp. 578–588, 2005.

... (1) so today it was an energized president **CLINTON** who formally presented his one point seven three trillion dollar budget to the congress and told them there'd be money left over first of the white house a.b.c's sam donaldson (2) ready this (3) morning here at the whitehouse and why not (4) next year's projected budget deficit zero where they've presidental shelf and tell *this* (5) *budget marks the hand of an era and ended decades of deficits that have shackled our economy paralyzed our politics and held our people back* ..... (6) [empty] (7) [empty] (8) administration officials say this balanced budget are the results of the president's sound policies he's critics say it's merely a matter of benefiting from a strong economy that other forces are driving for the matter why it couldn't come at a better time just another upward push for mr **CLINTON**'s new sudden sky high job approval rating peter thanks very ...

Fig. 1. Key-frames and corresponding speech transcripts for a sample sequence of shots for a story related to Clinton. Italic text shows Clinton's speech, and capitalized letters show when Clinton's name appears in the transcript. Note that, Clinton's name is mentioned when an anchor or reporter is speaking, but not when he is in the picture

is in the loop to select the relevant faces from a result set. However, in such systems, usually too many results are presented to the user making the retrieval process a highly time consuming task that is prone to errors.

In this study, we propose a method to extensively reduce the number of results for the user to examine. For this purpose, we propose a multi-modal approach and integrate the text and face information.

The success of using multiple modalities is shown for many multimedia applications [3]. Recently, similar attempts are made to search for people with the integration of text and face information [4, 5], and it is shown that such multimodal systems produce better performance than the text only based systems.

In this study, our goal is to reduce the number of results provided to the user by only taking the shots which are both aligned with the query names and include a face. Our proposed approach, first performs a text-based query and provides the shots aligned with the name of the person in the speech transcript and also the neighboring shots within a window. The results are pruned using a state-of-the-art face detection algorithm. The false faces produced by the face detector algorithm are further removed using a skin color based method.

Our main contribution is to use this pruned set to group the similar faces into some clusters, which are then used to generate some representative faces. Only, these representative faces which are on the order of tens are presented to the user. With the proposed approach the resulting set is extensively reduced, and therefore the search is made easy.

The experimental results will be presented on the six person queries of TRECVID 2004 evaluation: Bill Clinton, Saddam Hussein, Sam Donaldson, Boris Yeltsin, Benjamin Netanyahu, and Henry Hyde. The data set consists of 248 movies (30 minutes each) from ABC and CNN broadcast news.

2 Integrating Faces and Names

Generally, a user querying on a specific person wants to see the face of the person in the image and most probably prefers close-up views of the person. With this assumption, we incorporate the face information into a text based query system and find the faces associated with the query names. For this purpose, first the query names are searched over the speech transcripts which are aligned with the shots using the time information. Each shot is represented by a single key-frame and a face detection algorithm is applied on the key-frames of the shots associated with the query names to detect the faces. For this purpose, we have used a state-of-the-art face detector, which is Mikolajcyzk [6] implementation of Schneiderman-Kanade's face detection algorithm [7].

The Schneiderman-Kanade algorithm was reported to have an accuracy of 80.4% in Kodak test set for all faces [7]. However, the face detection performance of this algorithm on the TRECVID 2004 data set are observed to be much worse. Also, it is observed that the algorithm is less successful on detecting profile faces. These are mostly due to the great variation of pose and illumination in the data set and low resolution quality of the images.

For the rest of the results, we limit ourselves to the face detector output results, and provide only the shots which are both associated with the query names and/or surnames and include one or more faces as the results. Therefore the recall rate of the method is limited to what face detector extracts. However, the accuracy of the face detector is low and produce many false alarms. The time complexity of the retrieval system is high since the user has to search over a very large number of faces (on the order of hundreds and thousands). This process is also open to the errors, since the user can miss related faces among the many other unrelated ones. In order to overcome these problems and to increase the performance of the face detector by reducing the amount of false positives, we have applied a skin detector on the found face areas.

2.1 Improving Face Detection Accuracy Using Skin Color

Skin detection has been widely discussed in the literature and a recent survey on this topic is presented in [8]. Although Bayesian histogram method was claimed to achieve the highest accuracy in this review, in our preliminary experiments on two videos, we observed that the simple Gaussian probability distribution yielded better performance. Therefore, we modeled the probability of a pixel being a skin pixel, using Gaussian probability distributions on HSV color space, which is reported to be effective in discriminating skin pixels [8].

First, all images are converted to the HSV color space to determine their skin area. Using representative areas selected from 30 key-frames for skin and non-skin pixels, a unimodal Gaussian distribution is modeled. This model is then used to approximate the class-conditional probability of pixels to classify as skin or not-skin. That is, any given pixel is classified as a skin pixel if its Mahalanobis distance to skin model is less than a pre-defined threshold.

Fig. 2. Examples to the false detections eliminated by using **top:** the average skin color, **bottom:** the number of skin pixels

Table 1. Precision and recall values for three movies to compare the proposed skin color based methods with the original face detection

	Original	Average Skin Color	Number of Skin Pixels
Precision	0.41	0.71	0.77
Recall	0.40	0.38	0.38

We used the skin color detection method to eliminate the false alarms produced by the face detection algorithm. The lowering of the confidence level for the face detection algorithm increased the recall, but it also increased the number of wrong faces. In order to eliminate these wrong faces, we checked whether, (i) the average skin color value is less than a specified threshold value, and (ii) the number of skin pixels are fewer than a specified number. Figure 2 shows some of the faces eliminated. The overall increase in the detection performance is presented in Table 1. We would like to point out that there is a noticeable increase in precision along with a slight decrease in recall.

2.2 Retrieval Using the Combination of Text and Face Information

The retrieval on person queries are performed by first searching over the name of the person in the speech transcript and then applying a face detection algorithm to get only the shots including the name of the person and one or more faces. We call the method which prunes the result of text based query using only Schneiderman-Kanade's face detection algorithm as *text-and-face-based*, and the method which further eliminates the false faces by the skin color based method as *text-and-skin-based* method.

The comparison of these two methods based on the the number of correctly found faces over the number of all faces retrieved (called as *retrieval performance*) are given in Table 2. Note that more than one face can be detected in a single

Table 2. Number of faces correctly retrieved over total number of retrieved faces for each person query using two different methods

	Clinton	Saddam	Sam Donaldson	Yeltsin	Netanyahu	Henry Hyde
text-and-face-based	65/1113	8/127	36/114	8/69	4/35	1/3
text-and-skin-based	65/732	8/98	36/98	8/52	2/20	0/3

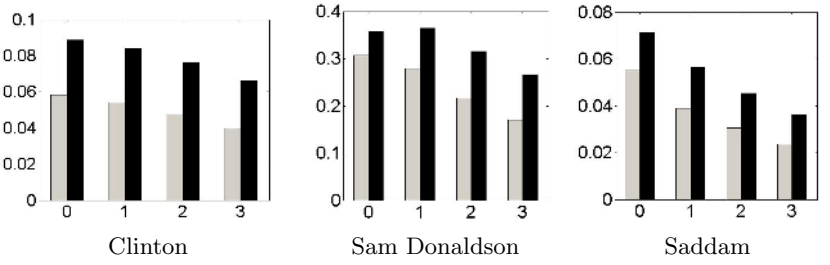


Fig. 3. Comparison of the retrieval performance when shots corresponding to the text are extended with the neighbors. **Gray:** when original face detection is used together with text, **black:** when skin color is used to improve the performance. Note that the scales are different. Maximum performances are 9% for Clinton, %36 for Sam Donaldson and %7 for Saddam queries

shot. In total, there are 1461 faces provided to the user with the *text-and-face-based* method, and 1003 faces with the *text-and-skin-based* methods. Among the final set of faces 122 faces are corresponding to the query people when only the face detection is used, and 119 faces are the correct faces when skin-based method is used. That is, the overall retrieval performance is 8% for *text-and-face-based* method, and increased to 12% with the *text-and-skin-based* method.

2.3 Extending to the Neighboring Shots

Due to the nature of the news videos, the name of a person is mostly mentioned when the anchor or reporter is speaking, whereas the face of the person actually appears a few shots before or later. Based on this observation, to find a person, instead of using only the shots where the name is mentioned, we also used the preceding and proceeding shots over a neighborhood. Specifically we experimented taking only the shot associated with the transcript on the time-basis (*Shot0*), and taking the N preceding and N following shots, where N is 1, 2 or 3 (*Shot1, Shot2* and *Shot3* respectively).

In Figure 3, the effect of taking the neighboring shots are shown using the text-and-face-based and text-and-skin-based methods for three of the queries. As can be observed from the figures, text-and-skin-based method always produces better results than text-and-face-based results. It is interesting to see that, taking the neighboring shots give worse performance than taking only a single shot when the simple integration of text and face information is used. As will be shown in the following sections, in some cases using the neighboring shots can produce better results when our proposed grouping method is used.

3 Grouping Similar Faces

We achieved an improvement in the retrieval performance with the proposed skin-based method, however the number of results presented to the user was still

high. The main reason for this is that the name of the query person is usually mentioned when an anchorperson or reporter is speaking. As a consequence of this, besides the faces of the queried person, many anchorperson or reporter faces along with faces of other unrelated persons are also extracted.

Let's consider a set of faces corresponding to the same person. Although different pose and illumination conditions will create different views of the same person, there will be also some similar conditions which result in similar views. Therefore, for a person, we expect that there will be a few number groups corresponding to different conditions and in each group there will be similar faces.

Based on this observation, we clustered the extracted faces into a number of groups. We assumed that, faces of the query person will be collected in a few groups and these will be different from the groups of the anchorperson or reporter faces. The other faces which appear only a few times will not create individual groups but will be distributed among the others according to their similarity.

3.1 Feature Extraction

In our experiments, we used three different features to represent face regions in vectoral form. The first one, called the color feature, consists of the mean and standard deviations of the 6×5 grid regions from the face image. The mean and standard deviations of 30 grids are computed in RGB form, resulting in 30x6=180 features for each face image. As the second feature set, we applied PCA to the images and took the first 40 dimensions as representative features. As the third feature set, we applied ICA on face images, with a learning rate of 0.5. The combination of PCA and color features and also ICA and color features are also constructed.

3.2 Clustering Strategy

The features extracted from face regions are used to cluster similar faces. Ideally, all faces of a particular person is collected in a single cluster. However, due to the unavailability of face-specific features, and noise, this is a seldom case. Therefore, we limited our goal as to cluster most of the images for a specific person only in a few groups and try to make these groups as coherent as possible.

One of the simplest algorithms for clustering is K-means. However, the choice of a constant K is by no means optimal. In this study, instead, we determine the number of clusters K adaptively using the G-means algorithm [9]. G-means clusters the data set starting from small number of clusters, K, and increases K iteratively if some of the current clusters fail the Gaussianity test (e.g., Kolmogorov-Smirov test).

In order to select the best feature for obtaining the best clusters, we perform a comparative experiment for the query on Clinton for Shot0. In Figure 4, number of target faces reached is plotted against the number of clusters target face is distributed to. For example, with color features, 90% of the correct faces are distributed into 8 clusters, whereas with PCA into 9 clusters, with PCA and color into 10 clusters, and with ICA into 22 clusters. According to these results,

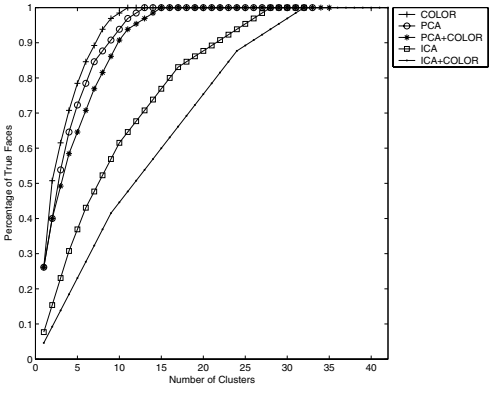


Fig. 4. Comparison of different methods on Clinton clusters

Fig. 5. Some clusters including Clinton faces. The percentage of Clinton faces over all of the faces are 43% for the first cluster, 94% for the second cluster, and 47% for the third cluster

Fig. 6. Selected examples from some anchor clusters. The true anchor occurrences are: 36/44 (82%) for the first cluster, 44/54 (82%) for the second cluster , 16/17 (94%) for the third cluster

color features give the best performance in this architecture since they collect the correct faces in least number of clusters. Similar patterns are observed for the other people and for the other shot windows. Therefore, in the rest of the experiments we apply color feature extraction to obtain the face groups.

In Figure 5 and Figure 6, example clusters including the faces of Clinton and some anchors are shown. As can be observed, the *coherence* which we define as the number of the most dominant face over all the faces in that cluster, is very high.

4 Retrieval Using Representative Faces

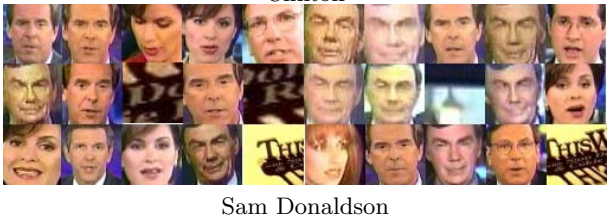
Since the clusters are sufficiently coherent, only a single face can represent the whole cluster. These faces, which we call as *representative faces*, can be selected as the ones closest to the mean of the cluster. Figure 7 shows the representatives for `Clinton` and `Sam Donaldson` queries.

We propose a retrieval strategy using these representatives to reduce the number of results presented to the user. The idea is that, when the groups are sufficiently coherent, then it is possible to represent them by a single face. The user is then inspect only these representatives instead of all of the faces.

In Table 3, the number of all representatives and the number of representatives corresponding to the query person are given. For example, for the Clinton query, only 24 representative faces are presented to the user and he/she selects 5 of them corresponding to Clinton faces. Note that, these numbers are only on the order of tens. This is a big reduction when compared to the initial results which are on the order of hundreds and thousands (see Table 2).

As can be also observed from Table 3, most of the faces of the query person resides in the selected clusters. For example, when Shot0 is considered, there are in total 65 Clinton faces in the data, and 51 of them resides in the selected clusters which have a Clinton face as a representative. Therefore, 78% of the correct faces can be retrieved by only viewing 24 faces, and selecting 5. This number is even higher for the other queries : 97% for Sam Donaldson and 100% for Saddam. We observe that, when the neighboring shots are considered, more noise is included in the data, therefore there is a decrease in the percentage of correctly retrieved faces.

Clinton



Sam Donaldson

Fig. 7. Representatives for Clinton and Sam Donaldson queries

Table 3. When the representatives corresponding to the query person are selected, number of clusters with the representative of the query person over the total number of clusters, and the number of correct faces in the selected clusters over the total number of correct faces are given. For example, consider Clinton when only a single shot where his name appears is taken (Shot0). (5/24)-(51/65) means that 24 clusters are obtained, and 5 of them have representatives with Clinton faces; also inside these 5 clusters, there are 51 Clinton faces, and the total number of Clintons in all the clusters is 65

	Shot0	Shot1	Shot2	Shot3
Clinton	(5/24)-(51/65)	(5/44)-(58/138)	(10/72)-(72/158)	(7/66)-(66/170)
Sam Donaldson	(9/30)-(35/36)	(8/30)-(76/89)	(8/26)-(98/106)	(8/26)-(101/114)
Saddam	(5/22)-(8/8)	(3/26)-(5/13)	(1/30)-(2/14)	(2/30)-(6/14)

Table 4. Retrieval performance when the clusters with the representatives of the query person are selected

	Shot0	Shot1	Shot2	Shot3
Clinton	40%	39%	43%	40%
Sam Donaldson	90%	81%	68%	61%
Saddam	80%	45%	100%	32%

In order to compare with the previous results, we use the retrieval performance and report the number of faces of the query person over all the faces in the selected clusters. The results are shown in Table 4. Since, the false alarms are also highly reduced, there is a big increase in the retrieval performance (compare with Figure 3).

We have also experimented *anchor filtering* which is previously proposed in other studies to improve the retrieval performance [5, 10]. The representative faces corresponding to anchors are selected and then these clusters are removed from the resulting set. The retrieval performance is evaluated on the remaining clusters. As it is shown in Table 5 and Table 6, the retrieval performance is worse than selecting the representatives corresponding to query people although almost all of the query faces can be found in the remaining clusters.

Table 5. When anchors are selected and removed, the number of clusters with the representative of the query person over the total number of clusters, and the number of correct faces in the selected clusters over the total number of correct faces are given

	Shot0	Shot1	Shot2	Shot3
Clinton	(8/24)-(64/65)	(13/44)-(136/138)	(18/72)-(155/158)	(15/66)-(168/170)
Sam Donaldson	(6/30)-(36/36)	(10/30)-(84/89)	(5/26)-(106/106)	(3/26)-(112/114)
Saddam	(5/22)-(8/8)	(6/26)-(12/13)	(5/30)-(13/14)	(6/30)-(13/14)

Table 6. Retrieval performance when the clusters of the anchor representatives are selected and removed

	Shot0	Shot1	Shot2	Shot3
Clinton	19%	14%	10%	10%
Sam Donaldson	56%	56%	39%	32%
Saddam	14%	8%	6%	5%

5 Discussion and Future Work

In this study, we propose a multi-modal approach for retrieving specific people from the news videos using both text and face information. Our main contribution is to extensively reduce the number of images provided to the user, and therefore increase the speed of the system with a large amount and at the same time not to loose many of the relevant images. For this purpose, the similar faces are clustered into groups, and representative faces are selected from each cluster to be provided to the user.

Similar clustering approach is proposed to name the faces in news photographs [11]. In their work, the images that contain a single face and a single name are used as a way of supervision to learn the name-face association. In our case, we usually do not have such a strong correspondence since in most of the times, when a single name is mentioned, the face corresponds to the anchorperson but not to the correct person. Similar approach can be adapted by manually choosing a set of correct faces and then using this information for supervision.

The success of the proposed method is limited by the accuracy of the initial face detection algorithm that we have used. We have noticed that almost half of the related shots are removed at the first step. Face detection algorithm should be improved not to miss any correct face. Also, the features that we have used are not face-specific. Better features should be studied in order to obtain more coherent clusters.

Acknowledgements

This work is supported by TÜBİTAK Career Grant 104E065 and Grant 104E077. We would like to thank Krystian Mikolajczyk for providing us the face detector code.

References

1. TREC Video Retrieval Evaluation http://www-nlpir.nist.gov/projects/trecvid/
2. Zhao, W. ,Chellappa, R., Phillips, P. J., Rosenfeld, A., "Face recognition: A literature survey", In ACM Computing Surveys, 2003.
3. Snoek, C.G.M., Worring, M., "Multimodal video indexing: A review of the state-of-the art", In Multimedia Tools and Applications, 25(1):5-35, January 2005.

4. Satoh, S., Kanade, T., "NAME-IT: Association of face and name in video", In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 1997.
5. Yang, J., Chen, M.-Y., Hauptmann, A., Finding Person X: Correlating Names with Visual Appearances Int'l Conf. on Image and Video Retrieval (CIVR), Ireland, July 21-23, 2004.
6. Mikolajczyk, K., "Face detector", Ph.D report, INRIA Rhone-Alpes.
7. Schneiderman. H., Kanade, T. "Object detection using statistics of parts", International Journal of Computer Vision, 2002.
8. Phung, S.L., Bouzerdoum, A., Chai, D., "Skin segmentation using color pixel classification: analysis and comparison", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 27, no.1, January 2005.
9. Hamerly, G., Elkan, C., "Learning the k in kmeans", Proc. of the NIPS 2003.
10. Duygulu, P., Hauptmann, A., "Whatś news, whatś not? Associating News videos with words" Int'l Conf. on Image and Video Retrieval (CIVR), Ireland, July 21-23, 2004.
11. Miller, T., Berg, A.C., Edwards,J., Maire, M., White, R.,Teh, Y.-W., Learned-Miller, E. Forsyth, D.A., "Faces and names in the news", In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2004.