Naming Faces in Broadcast News Video by Image Google

Chunxi Liu[1,3], Shuqiang Jiang[2], Qingming Huang[1,2,3]
[1]Graduate University of Chinese Academy of Sciences, Beijing, P. R. China, 100190
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, Beijing, P. R. China, 100190
[3]China-Singapore Institute of Digital Media
{cxliu, sqjiang, qmhuang}@jdl.ac.cn

ABSTRACT
Naming faces is important for news videos browsing and indexing. Although some research efforts have been contributed to it, they only use the concurrent information between the face and name or employ some clues as features and use simple heuristic method or machine learning approach to finish the task. They use little extra knowledge about the names and faces. Different from previous work, in this paper we present a novel approach to name the faces by exploring extra knowledge obtained from image google. The behind assumption is that the faces of those important persons will turn out many times in the web images and could be retrieved from image google easily. Firstly, faces are detected in the video frames; and the name entities of candidate persons are extracted from the textual information by automatic speech recognition and close caption detection. Then, these candidate person names are used as queries to find the name related person images through image google. After that, the retrieved result is analyzed and some typical faces are selected through feature density estimation. Finally, the detected faces in the news video are matched with the faces selected from the result returned by image google to label each face. Experimental results on MSNBC news and CNN news demonstrate that the proposed approach is effective.

Categories and Subject Descriptors: I.2.10 [**Vision and Scene Understanding**]: *Video Analysis*. H.3.3 [**Information Storage and Retrieval**]: *Information search and retrieval-search process*.

General Terms: Algorithms, Design, Experimentation

Keywords: News video analysis, naming faces, news video browsing and indexing.

1. INTRODUCTION
News video is an important media in our daily life. Every day a large volume of news video data is produced. However, without intelligent processing and understanding, it is difficult for people to access the interested topics, which scatter in the large accumulation of new video data. News video analysis has been a hot research topic for a long time in order to make people accessing them more easily. News mainly focuses on human activities and the faces of important peoples often appear in the news video tapes (some example faces are shown in Figure 1). If those faces in the news video could be automatically identified with their names, it will be helpful for news video indexing and retrieval and can facilitate news video personalization.

Figure 1. Example faces from news videos.

Existing work on naming faces can be generally classified into three classes: co-occurrence based method, matching based method and machine learning based method. For name face co-occurrence based method, the Name-It [1] system has been proposed. A face is labeled with the name that has the largest temporal overlap with a group of images containing faces that are similar to the given one. The face corresponding to a given name can be found in a similar way. This work is theoretically sound, while no serious performance evaluation has been reported. A similar work [2] labels face images in online news articles with names extracted from news captions by exploring co-occurrence information between the clustered faces and the names. For matching based method, a graph based approach [3] is proposed to find the most similar subset among the set of possible faces associated with the query name, where the most similar subset is thought to correspond to the faces of the queried person. The similarity of faces is represented by using SIFT descriptor. The interest points matching on two faces are decided under two constraints, namely the geometrical constraint and the unique match constraint. The most similar set of faces is then found based on a greedy densest component algorithm. The result is fine but is not as good as the supervised method. For machine learning based method, Yang et al. [4] formulated the person naming problem into a learning framework, which predicts the most likely name for each person based on the features and refines the predictions using the constraints. The features help distinguishing the true name of every person and the constraints reveal the relationships among the names of different persons. Further, in order to overcome the problem of large training data requirement, Yang et al. [5] proposed to use the multiple instances learning algorithm to label the faces in the broadcast video. Although both results are good, they all need labeling training data and better result can be performed by using more labeled data.

Although many approaches have been proposed for naming faces, most of them use the information contained in the data and little extra information is used. In this paper, we propose a novel framework for naming faces in news video by utilizing the extra knowledge obtained from image google. We formulate naming faces into a constrained face recognition problem and the face

MM'08, October 26-31, 2008, Vancouver, British Columbia, Canada.

recognition database is built dynamically through image google. The comparison of different naming faces strategy is shown in Figure 2.

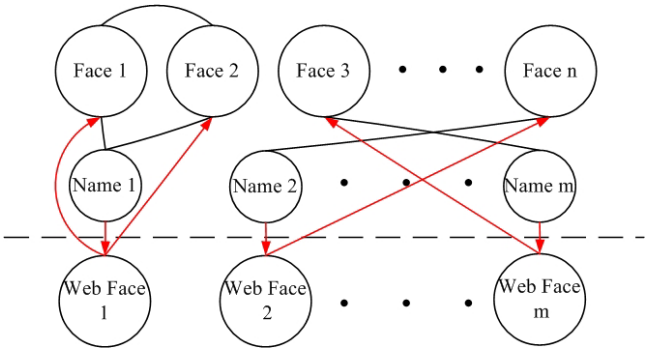

Figure 2. Naming faces strategy comparison.

Most existing work only use the information contained in the video data and employs features between face and name and features between face and face, which are shown upper the line in Figure 2. While in our approach, except these features in the video, extra web face features, which are shown under the line in Figure 2, are used to help label each face.

The rest of the paper is organized as follows. Section 2 gives an overview of the system. Section 3 describes the news structure analysis and face detection. Section 4 describes name entity extraction and image retrieval from image google. The naming faces is achieved in section 5. Finally, the experimental results are given in section 6.

2. SYSTEM OVERVIEW

The proposed framework of our approach is shown in Figure 3.

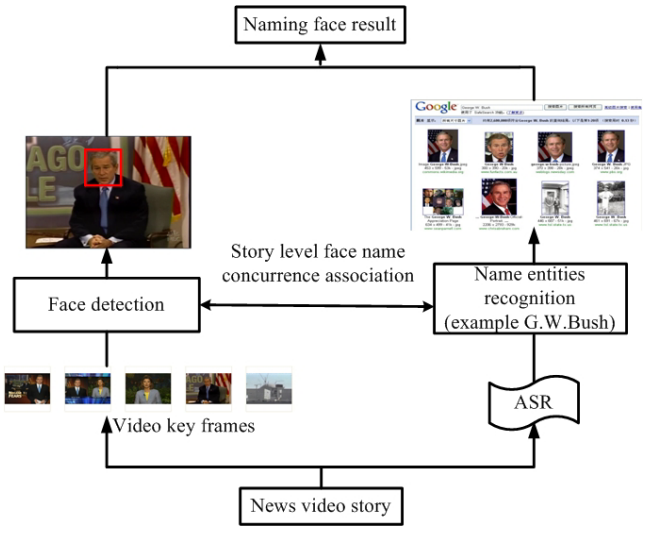

Figure 3. The framework of naming faces for news video.

Our approach utilizes the story level co-occurrence constraint between the name and face in news video. That means only the face and name turn out in the same story will be analyzed. Firstly, the news video stories are obtained. The faces in the key frames and the names in the automatic speech recognition result (ASR) and close caption are extracted. Then, based on the temporal constraint of story, the face and name associations are built. After that, the extracted names are used as queries to retrieve images from image google. Finally, the retrieved faces are selected through feature density estimation and matched with the faces in the video to help label them. The main contribution of the paper lies in that we propose a novel framework for naming faces by exploring extra knowledge from image google. This method is straightforward and does not need labeling training data.

3. NEWS VIDEO STRUCTURE ANALYSIS AND FACE DETECTION

In our approach, the name and face association in the video is constrained by the story, which means that the associated name and face are in the same story. Then in order to identify the faces with the names, the first step is to obtain the distinct stories. In order to segment news video into stories, the state of the art approach is adopted here [6]. By using the trained HMM model and the viterb algorithm, the stories are obtained. News video consists of many different stories and also includes weather reports and advertisement *etc*. We focus on labeling the faces appear in the formal news stories especially stories reporting important international affairs. Therefore, as a preprocessing step, the advertisement and weather report are removed.

For naming faces, the faces should be obtained firstly. Face detection is a long-standing problem and many approaches have been proposed. We adopted the method proposed by [7]. This method can achieve good performance and more details of the algorithm can be found in [7]. In our approach, each obtained story is further divided into shots and each shot is represented by the frame in the middle of the shot. Then face detection is performed on each key frame.

4. NAME ENTITY EXTRACTION AND FACE RETRIEVAL FROM GOOGLE

The name entity is important for naming faces in the video. In our approach, the name entities are extracted from the ASR data and the close caption. For the ASR data, the person names are extracted using an automatic name entity recognition tool [8]. By our observation, when a person is giving a talk in the video, his name is often labeled as close caption; two examples are shown in Figure 4.

Figure 4. Two examples of name labeled in the image.

From Figure 4, it can be observed that sometimes the name in the close caption is accordant with the face in the frame (the left one) and sometimes is not (the right one). Even though, the names in the close caption are important clues for naming faces in the news video. In our approach, the close caption is detected by a scene text detection algorithm [9] and the text is recognized by OCR software [10].

By our observation, a person face may be shared by different person names. This is due to the fact that sometimes the person is referred by the last name and sometimes by the full name *etc*. Take "*George W. Bush*" as an example, the often used names are "*Bush*", "*George Bush*", "*George W. Bush*", "President *George W. Bush*", "President *Bush*", *etc*. We calculate the frequency of different parts of the name entity from google, and the normalized probability distribution of them is shown in Figure 5.

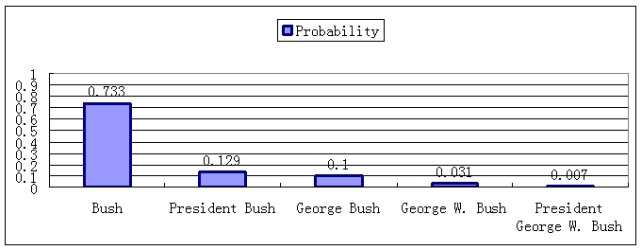

Figure 5. Normalized name probability distribution.

From Figure 5, we can see that with the name becoming more specific, its turn out probability in the WWW becomes less. In our approach, in order to retrieve the true person faces from image google, the name with strongest constraint is used, which will be the longest one. One important issue here is how to map the different names of the same person into one. From the example above we can see that the last name of the person appears in every name example. Therefore, we achieve the goal of mapping names by matching the last name. For the above example, *President George W. Bush* will be used as the final name query to retrieve the name related person images from image google.

Figure 6. The top 20 images retrieved for query *President George W. Bush.*

The retrieved result from image google always contains tens of thousands of images. The image google indexing is dynamic and when we perform the above *bush* example totally 824,000 images are retrieved. Processing all of those images is time consuming and is not necessary for our application. On the other hand, the top images are the most confident ones retrieved by the search engine. Thus, we only consider the retrieved top 20 images. The top 20 images for *President George W. Bush* are shown in Figure 6.

In Figure 6, some of the faces look like very small, but actually they are big enough in the original images. We can also see that although most of the retrieved top 20 images contain the wanted frontal person face, some retrieved images are of no help for our naming faces (e.g. image 12, 13, 16, 17, 19). Thus, we have to select proper images from the retrieved result. For the top 20 images, face detection is first performed to obtain the faces in each image. If there are no faces or more than one face in the image, the image will be removed. There are also some cases that the retrieved face is not the true one. Therefore, we have to properly select the right faces. In our approach, the faces are selected based on the assumption that most of the retrieved faces are right ones and the similarity between the faces of the same person is higher than the similarity between the faces of different persons. Generally, since all the relevant images are related to the target person, the feature vectors corresponding to them should be relatively close in the feature space; while the feature vectors corresponding to the irrelevant person faces should be relatively scattered since they are related to different persons. Thus, looking for high density region in the face feature space could be helpful to identify name related person faces.

Let $C=\{x_1,x_2,..,x_{|C|}\}$ denote the set of top-ranked candidate faces. Each is represented by a d-dimensional feature vector, which will be described in section 5, and z denotes a point in the feature space. We employ a simple method to estimate the density [11], as shown in Equation 1.

$$Density(z)=\sum_{i=1}^{|c|} e^{-\sum_{j=1}^{d}|z_j-x_{ij}|^2} \qquad (1)$$

In our approach each face is estimated by equation 1. Then, the faces are sorted according to their densities and the top 5 (or 3 if the number of frontal face is small that 8) candidate faces with higher densities are deemed as the typical query related faces. The selected five face images for the above *bush* example are shown in Figure 7 in descending order according to their densities.

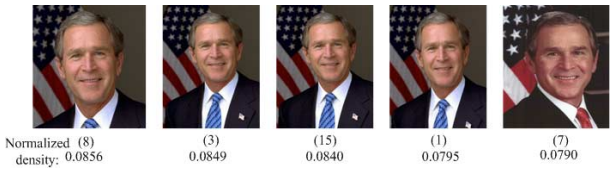

Figure 7. The selected top 5 images for *bush*.

5. FACE MATCHING

After obtaining the 5 typical faces from web images, the next step is to use them to confirm whether the face in the video is the name related one. By now, we can see that our naming faces problem is a constrained face recognition problem. The recognition face database could be built dynamically through search engine. Our naming faces confirmation scheme is shown in Figure 8. The final label is decided by averaging the matching result of the five faces.

Two important issues for face recognition are how to represent each face and how to match the faces. In our approach we adopt the local Gabor feature recognition approach in [12]. Before feature extraction, the size of each face image is normalized to 128 by 160 pixels with the eye distance being 72 pixels. Totally 40 Gabor wavelets are used and the final feature dimension is

40960. In order to use Fisher Discriminate Analysis to deal with the problem, these features are down-sampled and the dimension is reduced from 40960 to 640. For face recognition a similarity threshold should be set and in our approach the threshold is 0.55. The similarity between the face from video and web is decided by averaging the similarities of the video face with the five selected faces. If there is more than one name matched with the face, the face will be labeled as the name with most similarity faces.

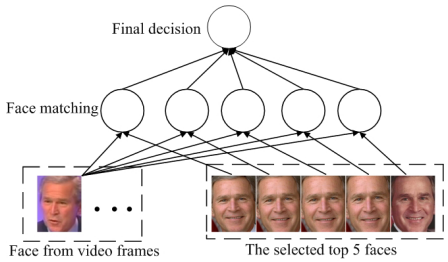

Figure 8. Face matching scheme.

6. EXPERIMENTS

We apply the proposed method to associate name and face in 5 broadcast videos, 3 CNN news video and 2 MSNBC news video. They are selected from the TRECVID 2005 data and the resolution is 352x240. We focus on analyzing those important international affairs and the local news and entertainment news will be not considered. Totally, 139 faces and 86 names are extracted from the selected news video stories.

In order to see if the image google can provide good background for our naming faces application, we use the extracted names as query to search in the google image search engine. We count the number of retrieved frontal faces and the number of right retrieved faces. The result shows that the mean average precision for frontal face is 0.53, and for the retrieved faces the right retrieved face precision is 84.6%. This is means that the mean average precision for the right retrieved face is 0.449. Finally, we test the performance on the video data and the result is shown in table 1.

Tabel 1. Naming faces experimental results

Correct named face	Wrong named face	Correct discriminated face with no name
91	16	32

From table 1 we can see that our naming faces performance is promising and is better than the result reported in [3][5]. There are also several factors affecting our naming faces performance. Firstly, sometimes the face resolution in the video is limited. Secondly, the retrieved images may contain no right faces (this happens especially when the person is not so famous). Thirdly, there could be many faces corresponding to the queried person in different conditions, poses and times. Face recognition is a long standing and well studied problem. For larger and more realistic data sets, face recognition is difficult and error-prone due to large variations in pose, illumination and facial expression.

7. CONCLUSION

Naming faces is important for news videos browsing and indexing. In this paper we present a novel approach to name the faces by exploring extra knowledge obtained from image google. We formulate the naming faces into a constrained face recognition problem. The assumption behind is that the faces of the important persons will turn out many times in the web image and can be retrieved through image google. Name entities extracted from the automatic speech recognition text are used as queries to search the names related face images from image google to build the face recognition database. The final naming faces decision is decided by face matching. The experimental results on MSNBC news and CNN news demonstrate that the proposed approach is effective. In the future we may use other method such as local interest points to help match faces.

8. ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under Grant 60702035 and 60773136, National Hi-Tech Development Program (863 Program) of China under Grant 2006AA01Z117 and 2006AA010105, and Science100 Plan of Chinese Academy of Sciences: 99T3002T03.

9. REFERENCES

[1] S. Satoh, T. Kanade. NAME-IT: Association of Faces and Names in Video. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp368-373, 1997.

[2] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.W. Teh, E. Miller, D. A. Foryth. Names and Faces in the News. In *Proc.of IEEE Conf. on Computer Vision and Pattern Recognition*, pp848-854, 2004.

[3] D. Ozkan and P. Duygulu. A Graph Based Approach for Naming Faces in News Photos. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp1477-1482, 2006.

[4] J. Yang and A. G. Hauptmann. Naming Every Individual in News Video Monologues. In *Proc. of ACM Int'l Conf. on Multimedi*a, pp580-587, 2004.

[5] J. Yang, R. Yang, and A. G. Hauptmann. Multiple Instance Learning for Labeling Faces in Broadcasting News Video. In *Proc. of ACM Int'l Conf. on Multimedi*a, pp31-40, 2005.

[6] L. Chaisorn, T.-S Chua, C.-K Koh, Y.-L Zhao, H. Xu, H. Feng and Q. Tian. A two-level Multi-modal Approach for Story Segmentation of Large News Video Corpus. *TRECVID workshop*, 2003.

[7] J. Chen, X. Chen, W. Gao. Expand Training Set for Face Detection by GA Re-sampling. In *Proc. IEEE int'l conf. on automatic face and gesture recognition*, 2004.

[8] Alias-i. Lingpipe named entity tagger. In http://www.aliasi.com/lingpipe/.

[9] Q. Ye and Q. Huang. A New Text Detection Algorithm in Image/Video Frames. *Lecture note in computer science*, *Pacific-Rim conference on Multimedia*, pp858-865, 2004.

[10] http://www.hw99.com

[11] K. Fukunaga and L. Hostetler. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. *IEEE Transactions on Information Theory*, vol.21, no.1, pp32–40, 1975.

[12] Y. Su, S. Shan, X. Chen and W. Gao. Hierarchical Ensemble of Global and Local Classifiers for Face Recognition. In *Proc. of IEEE Int'l Conf. on Computer Vision*, pp1-8, 2007.