# Automatic Naming of Speakers in Video via Name-Face Mapping

Zhixin Liu[1], Cheng Jin[1], Yuejie Zhang[1], Tao Zhang[2]

[1]School of Computer Science
Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai 200433, P. R. China
[2]School of Information Management & Engineering,
Shanghai University of Finance & Economics, Shanghai 200433, P. R. China
{14210240048, jc, yjzhang}@fudan.edu.cn, taozhang@mail.shfeu.edu.cn

**Abstract.** The problem of automatically labelling the appearances of characters in video with their names is challenging due to the huge variation in the appearance of each character and the weakness and ambiguity of available annotations. We can achieve high precision by combining multiple sources of information, both visual and textual. The principal novelties that we introduce in this paper are: (i) extracting face features in video by neural network; (ii) strengthening the mapping between names and faces by analyzing the co-occurrence of names and faces; (iii) automatically and efficiently labelling appearances of main characters with their names.

**Keywords:** Automatic Naming Speakers in Video, Face and Clustering Extraction, Name Extraction, Name-Face Mapping.

## 1 Introduction

With the rapid development of information technology, more and more multimedia data appears on the Internet, and there has been an explosive growth of audio-visual content. Content-based multimedia retrieval research develops rapidly. However, there are many limitations in multimedia retrieval, such as most retrieval manners are based on a single type of multimedia data, or the other modal data only plays a secondary role. To solve this problem, some researchers are interested in cross-media retrieval, namely matching multimodal information by a certain correlation to realize flexibly crossing different medium for retrieval. However, this often requires expensive manual annotations, especially for video contents.

Manual annotation of each new video source is expensive and impossible. An interesting alternative is using unsupervised approaches to name people in multimedia documents. In most previous works, the researchers first automatically classified each speech with an anonymous label and then used other methods to find the name of each class. Most previous works concerned the naming of people in video, and essentially

adfa, p. 1, 2011.

use the same framework: a) face clustering; b) extracting names for each person; and c) names/faces mapping. Such methods are different in how to cluster faces, how to extract names and how to match names with faces.

However, extracting names for each person in video is a very difficult problem, as subtitles usually do not directly describe the faces in video. Moreover, even if the name of a face is mentioned in subtitles, the alignment may be wrong. There are many other problems, such as there are often many faces in the same frame, and many names in the transcripts, or many unnamed faces or names that are mentioned but not displayed. Another difficult problem is that the identification of the person can be very hard due to the changes in pose, lighting conditions, facial expressions and partial occlusion. Recently, there has been a surge of interest in neural networks. In particular, deep and large networks have exhibited impressive results. However, most previous works on the recognition of characters in video did not use the novel technology, they still used the traditional method of clustering and traditional features in the face clustering stage. Another problem is that many previous works needed a lot of manual annotation information or did not make full use of the information in the video frame.

Based on these observations above, a novel scheme is proposed in this paper for facilitating more effective people news annotation via name-face Mapping by integrating multimodal information involved in video news. Our proposed scheme differs from other earlier work in multiple aspects, as shown in Fig. 1. a) In the face clustering stage, we mainly use the popularity technology of neural network to train a generic face classifier, and extract deep descriptors of the human face by the classifier. b) In the name extraction stage, not only the text information such as video introduction, subtitles are used, but we also use the OCR technology to extract the text information in each frame of video. All the information will be recorded on the video track, and will also be preliminary matched with the face in the first stage. The initial matching information can be found in this way. c) In the matching stage, an efficient optimization algorithm based on the fuzzy clustering is particularly established to verify the feasibility of our automatic name-face Mapping algorithm.

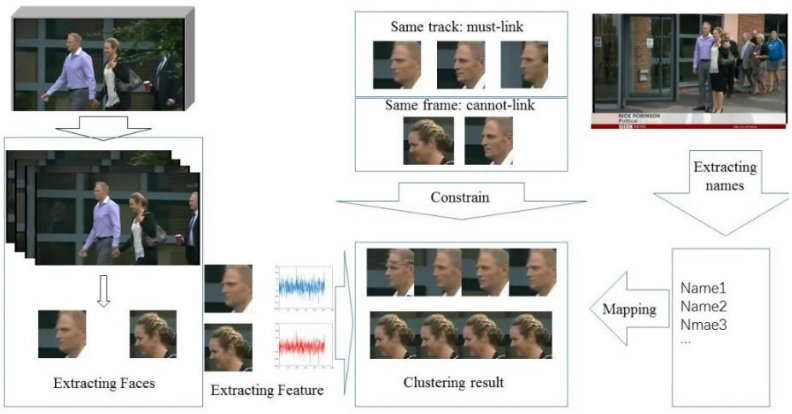


**Fig. 1.** The framework of automatic naming of speakers in video via name-face Mapping.

## 2 Related Work

Previous works on the naming of people in video, used essentially the same framework: a) Face clustering; b) Extracting names for each person; c) Names/Faces mapping. However, previous works on the recognition of characters in video had often ignored the availability of textual information. Many researchers obtained the names by manually annotation. A semi-automatic method was proposed in [1] to name face images in BBC news. It needed to manually name some faces and then used the iterative label propagation in a graph of connected faces or name-face pairs. In order to reduce the manual work, some researchers dug the information in the subtitles or audio. Everingham *et al*. in [2] obtained the high precision by combining multiple sources of information including subtitles, transcripts and visual information. Poignant *et al*. in [3] even tried to use the audio information to extract names. Besides extracting names, it was a challenging task to obtain clusters for per character without merging multiple characters into a single cluster. Zhou *et al*. in [4] made use of must-link and cannot-link constraints to cluster per character. However, it needed to know the number of clusters in advance and was difficult to achieve good results for videos.

Usually, we cannot obtain the high precision in face clustering by using the traditional representation of facial features. Recently, there has been a surge of interest in neural networks. In particular, deep and large networks have exhibited impressive results [5]. However, most previous works on the recognition of characters in video do not use such novel technology. Many other previous works needed a large number of manual annotation information or did not make full use of the information in the video frame. Many clustering methods neglected the particularity of video news.

We can obtain the high precision by combining multiple sources of information, both visual and textual. The principal novelties that we introduce are: (i) extracting face features in video by using the neural network; (ii) strengthening the mapping between names and faces by analyzing the co-occurrence of names and faces; (iii) automatically and efficiently labelling the appearances of main characters with their names.

## 3 Face Extraction and Clustering

This section describes how we extract faces and cluster persons. It aims to extract the faces in the video and extract the descriptors of their appearances. The descriptors can be used to match the same person and improve the final experimental results.

### 3.1 Face Detection

Many tools and algorithms can extract the face in the picture easily and quickly. In a 30-minutes video, we can extract tens of thousands of faces which belong to dozens of individuals.

Firstly, we need to extract the faces in the video. A video is composed of many coherent pictures and the technology of extracting the face region from the static image is very mature at present. In this paper, we use the V-J video face recognition method

for the extraction work. If we just simply extract the faces from each frame in video, we will get very large and complex human faces, and it will aggravate the burden of clustering. Fortunately, the frames in a video are not independent. There is strong continuity between frames. We can use this peculiarity to get a preliminary clustering of persons in video, which will greatly reduce the initial complexity of faces, and will improve the efficiency and accuracy of clustering results.

Face tracking is similar to object tracking. There are many ways to achieve good results at present. In this paper, we mainly use the optical flow method (Kanade - Lucas – Tomasi, KLT) [2] to track faces. The instantiation for the algorithm of face tracking is shown in the Fig. 2.

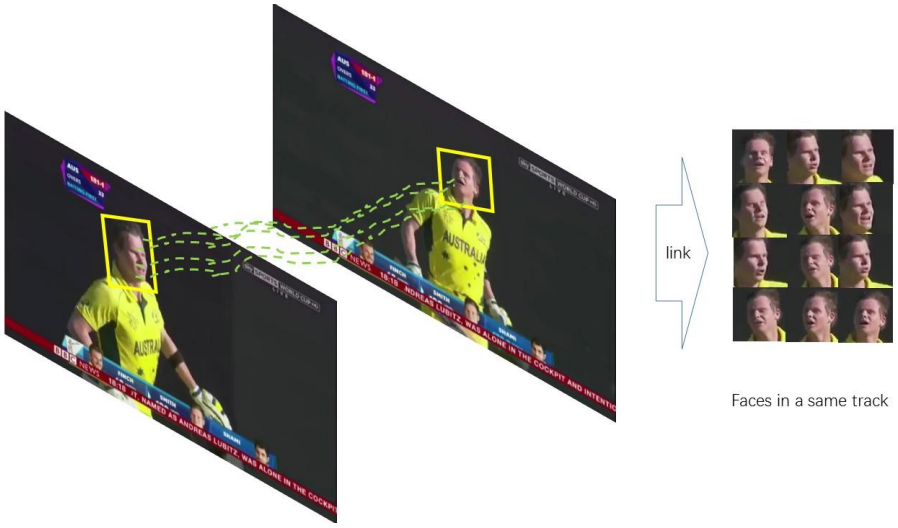


**Fig. 2.** An instantiation of face tracking by point tracking.

The KLT algorithm mainly tracks multiple points to track objects. In terms of face tracking, we can identify the feature points in the face region, then use the KLT algorithm to track such feature points in order to achieve the goal of tracking faces. The KLT algorithm can only track objects with small changes. Thus it can only be used in the continuous subset of video. We need to divide the video into smaller pieces according to the continuity. Considering the accuracy and efficiency, the segmentation criteria is the color histogram between two adjacent frames. We cannot achieve absolutely accurate identification of video segments by using the color histogram, but small mistakes just increase the burden of initial clustering and cannot lead to too much influence on the final result.

### 3.2 Feature Extraction

After the treatment in the previous section, we have already extract the faces of video, and carry on the preliminary clustering. However, because of the particularity of

video, the same person could still be divided into dozens to hundreds of different classes after the preliminary clustering. It is an almost impossible task to name all the classes, thus we need further clustering results.

In the previous work, most people used the traditional method to extract the features of each face, and designed a certain characteristic to cluster. The traditional methods generally are difficult to adapt to the complexity of faces in video. There is a huge bottleneck in traditional methods. Fortunately, we have a more accurate method of clustering with the emergence of neural network. In this paper, we use VGG Face Descriptor [6] to extract the features of faces. The CNN architecture A is given in full detail in Table 1. It can achieve the high precision in LFW and Youtube Faces Dataset.

**Table 1.** Network configuration.

| **Layer** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Type** | **In** | **conv** | **relu** | **conv** | **relu** | **pool** | **conv** | **relu** | **conv** | **relu** | **pool** | **conv** | **relu** | **conv** |
| **support** | — | 3 | 1 | 3 | 1 | 2 | 3 | 1 | 3 | 1 | 2 | 3 | 1 | 3 |
| **filt dim** | — | 3 | — | 64 | — | — | 64 | — | 128 | — | — | 128 | — | 256 |
| **num ilts** | — | 64 | — | 64 | — | — | 128 | — | 128 | — | — | 256 | — | 256 |
| **stride** | — | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| **pad** | — | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| **Layer** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** | **23** | **24** | **25** | **26** | **27** |
| **Type** | **relu** | **conv** | **relu** | **pool** | **conv** | **relu** | **conv** | **relu** | **conv** | **relu** | **pool** | **conv** | **relu** | **conv** |
| **support** | 1 | 3 | 1 | 2 | 3 | 1 | 3 | 1 | 3 | 1 | 2 | 3 | 1 | 3 |
| **filt dim** | — | 256 | — | — | 256 | — | 512 | — | 512 | — | — | 512 | — | 512 |
| **num ilts** | — | 256 | — | — | 512 | — | 512 | — | 512 | — | — | 512 | — | 512 |
| **stride** | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| **pad** | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| **Layer** | **28** | **29** | **30** | **31** | **32** | **33** | **34** | **35** | **36** | **37** | | | | |
| **Type** | **relu** | **conv** | **relu** | **pool** | **conv** | **relu** | **conv** | **relu** | **conv** | **softmax** | | | | |
| **support** | 1 | 3 | 1 | 2 | 7 | 1 | 1 | 1 | 1 | 1 | | | | |
| **filt dim** | — | 512 | — | — | 512 | — | 4096 | — | 4096 | — | | | | |
| **num ilts** | — | 512 | — | — | 4096 | — | 4096 | — | 2622 | — | | | | |
| **stride** | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | | | | |
| **pad** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | |

We align the faces before extracting the features and we use the approach in [7] to align the faces in this paper.

### 3.3 Face Clustering

We can obtain good clustering results by the excellent face features extracted from neural network and the inherent constraint in video: faces in same frame cannot be linked and faces in same track must be linked. In this paper, we cluster the faces by the distance proposed in [8] and the inherent constraint in video.

## 4 Name Extraction

In this section, we will mainly introduce how to deal with another important information in the cross-modal manner. The source of modern videos is variety, and a video is not often with enough text introduction, especially lack of the introduction of important characters in video. Finding the personal information of video is a difficult task in this field. Our treatment still cannot adapt to all of the videos, but can reduce the requirement of video data, especially the neural network technology used in OCR [9].

### 4.1 OCR Text Extraction

The script of OCR technology research has made remarkable achievements. We can say that the current print OCR recognition technology has reached a higher level. The OCR technology can be used to fully extract the textual information in video. By using OCR, we can extract the text information in each frame of video.

### 4.2 Name Identification

Although, we can easily get a lot of the text information, most of the test information is unserviceable. Only the person's name is useful. In this paper, we use the stanford-NER tool to extract the name in the text information. In order to cooperate with the mapping work, each name in the frame will be converted to the track of the name. We need to determine the name extracted by OCR not only the appearance on which frames, but also the position in each frame. Such information will enhance the mapping results.

### 4.3 Integration of Text

The previous processing fully extracts the names in video, but it also produces a lot of noises, such as geographical name, company name and the name in the scroll bar. We can see in the Fig. 3. that a lot of text information could be recognized as the name for the person in the same frame. We first eliminate the region with moving text and then eliminate the geographical names and company names. At last, we eliminate the very rare name.


**Fig. 3.** The interference information in the red region and the correct name in green region.

Now, we can easily get the correct text information by the above process. However, there are often many persons in the same frame. We need to find the most possible corresponding relationship between the name and the face to enhance the mapping results. We propose the following formula to determine the initial matching distance for the co-occurrence name and face:

$$RS(Name_i, Face_j) = \frac{\min(dist(Name_i, Face_j) * dist(Face_j, Frame\_center))}{area(Face_j)} \quad (1)$$

where $Name_i$ denotes the track of the $i^{th}$ name; $Face_j$ denotes the track of the $j^{th}$ face; min($dist$($Name_i$, $Face_j$)) denotes the min distance between $Name_i$ and $Face_j$; $Frame\_center$ denotes the center coordinate of the frame; area($Face_j$) denotes the sum area of the faces in the $i^{th}$ face track. For most of news videos, the name of a face often appears with the face at the same time and the face locates in the center of the frame. We propose the formula to reflect the regular pattern, as shown in the Fig. 4.


**Fig. 4.** The correct faces often appear in the middle region and are bigger than other faces.

## 5 Names-Face Mapping

We have the movement track of names and faces by the above processing. We will obtain eventually matching results through the analysis for the movement track of names and faces. We get the final match results through a matrix of names and faces.

Name-face mapping aims at finding the optimal one-to-one name-face matching in video. Some probability-statistical models have been used to solve this task. However, most of the strategies are just the process of clustering without mapping and need a lot of manual work [10]. Thus an improved Fuzzy C-Means (FCM) clustering algorithm, which introduces the consideration of salient name information and integrates the limitation of name-face co-occurrence, is established to make a better solution for name-face mapping. Compared to the general FCM clustering, the improved one can better describe the multimodal features in multimodal videos, and makes more reasonable solution for such a specific mapping issue.

We will use the following symbols:

*FS*: faces in the given video set;

*NS*: names in the given video set;

*FN*: the number of face feature vectors in the face set *FS*;

*NC*: name clusters in the name set *NS*;

*NCN*: the number of name clusters in the name set *NS*;

$Face_i$: the face feature vector for the $i^{th}$ face in *FS* and $1 \le i \le FN$;

$NC\_Center_j$: the prototype of the center of face cluster for the $j^{th}$ *NC* in *NS*, $1 \le j \le NCN$;

$U_{ij}$: the membership degree of $Face_i$ in $NC\_Center_j$;

*m*: the weighting exponent on each fuzzy membership degree, which is set empirically (usually set as 2).

Our improved FCM clustering algorithm mainly aims at optimizing the following objective function $J_m$ for every video.

$$J_m = \sum_{i=1}^{FN} \sum_{P_{ij}=1} U_{ij}^m * Dist(Face_i, NC\_Center_j) \tag{2}$$

where $Dist(Face_i, NC\_Center_j)$ is an Euclidean distance measure between $Face_i$ and $NC\_Center_j$; and $P_{ij}$ denotes that if the $i^{th}$ face and the $j^{th}$ name co-occur in the same time, $P_{ij}$ is set as 1, otherwise as 0. This function focuses on making the optimization for the distance among the face clusters associated with different *NC*s, so that each cluster has both the higher intra-cluster cohesion and the farther inter-cluster distance.

### 5.1 Initialization

We get the final matching results through optimizing a correlation matrix of names and faces. The initial value of the matrix has great influence on the final results. We need good initial values through the track of names and faces.

The track of faces is obtained by face detection, and the track of names is obtained by OCR. The track of names obtained by OCR is very important. The names in frames generally appear only once, and are often associated with the location of the relevant people. Thus we determine the initial value of the matrix by analyzing the matching degree of the track of names and faces in frames, which follow the following steps:

Firstly, for every trajectory of name, find all the trajectories of faces around the track of names. Secondly, for each finding trajectories of faces, calculate the minimum distance between the trajectories of faces and names.

Finally, the results of the divisor between the minimum distance of the face and the total trajectory distance is the initial value of Formula (1).

If the name extracted by the subtitle appearing in a frame, we will not consider the effect of the name. When the subtitle-name appears, the related face may appear at the same time, and also is likely to appear before or later. Subtitle-name tends to appear many times, thus we initialize the matching matrix by analyzing the co-occurrence of subtitle-name and the faces around it.

The initialization for *U* by using the name salience can be defined as:

$$U_{i,j} = \begin{cases} 0, & P_{i,j} = 0 \\ RS(Name_j, Face_i), & P_{i,j} \neq 0 \end{cases} \tag{3}$$

where $RS(Name_j)$ denotes the important degree of $Name_j$ that co-occurs with $Face_i$ in the same video; and because $P_{ij} \neq 0$, it can be sure that $Face_i$ and $Name_j$ exactly co-occur

in the video. Meanwhile, for the different names in a video, the sum of all their salience values is 1. The center of face cluster for each *NC* in *NS* is initialized according to Formula (4) with the parameter $P_{ij}$.

$$NC\_Center_j = \frac{\sum_{i=1}^{FN}\sum_{P_{ij}=1} U_{ij}^m * Face_i}{\sum_{i=1}^{FN}\sum_{P_{ij}=1} U_{ij}^m} \tag{4}$$

### 5.2 Matrix Iteration

There will be a lot of noises in the matching matrix because of the complexity and particularity of the video, and there will also be many correct initial values in the matrix matching. We should implement the process of iteration. This process aims at constantly amending the center of face cluster for each *NC* in *NS* according to Formulae (4) and (5), and also introduces the parameter $P_{ij}$ that plays an important role for controlling the iteration.

$$U_{ij} = \begin{cases} 0, & P_{ij} = 0 \\ \dfrac{1}{\sum_{P_{ik}=1}\left(\dfrac{Dist(Face_i, NC\_Center_j)}{Dist(Face_i, NC\_Center_k)}\right)^{\frac{2}{m-1}}}, & P_{ij} \neq 0 \end{cases} \tag{5}$$
$$1 \le k \le NCN$$

In every iteration process, the center of face cluster for each *NC* in *NS* and the membership degree of each face in the face cluster for each *NC* are both recalculated and updated. After such a process, the center of each face cluster and the membership degree of each face will become more precise.

### 5.3 Mapping

The above iteration process will stop until the center of face cluster for each *NC* in *NS* no longer has offsets, or the number of iteration times reaches the preset maximal value. With the iteration to a convergence state, the fuzzy partition matrix *U* and each center of face cluster are both output and taken as the final name-face association mapping results, as shown in Formula (6).

$$Name - Face\_Mapping(Face_i) = \arg\max_{NC_j}(U_{ij}) \tag{6}$$

## 6 Experimental Analysis and Discussion

### 6.1 Dataset and Evaluation Metrics

Our dataset is established based on the video website of *YouTube*! *News Data* constructed by ourselves, in which there are 11 news videos with around 200 names and about 40,000 faces. The proposed method is applied to the dataset (in total around 8

hours of video). The ground-truth names for every person are produced by manual annotation. To evaluate the effectiveness of our algorithm, Correct Mapping Rate (CMR) is defined as the percentage of correct mappings generated in the ground-truth.

**Table 2.** The statistical information about our Dataset.

| Total Faces | Clusterings | Clusterings (Name) | Clusterings (Null) | Clusterings (Unique Name) |
|---|---|---|---|---|
| **41,191** | **1,136** | **213** | **923** | **173** |

## 6.2 Experimental Results

To prove the correctness of the parameters we choose, we analysis the experimental results under different parameters. Table 3 shows the experimental results. The *co-occurrence* means the co-occurrence between the names and faces. The *dist*(*F*, *N*) means the min distance between the names and faces. The *dist*(*F*, *center*) means the average distance between faces and the center of the frame. The *area*(*F*) means the whole area of the faces.

**Table 3.** Experimental results under different parameters.

| Approach | *CAR* | | |
|---|---|---|---|
| | ***Faces with name*** | ***Faces without name*** | ***All faces*** |
| **co-occurrence** | 79.2% | 89.1% | 85.5% |
| **co-occurrence+dist(F,N)** | 81.3% | 89.1% | 86.2% |
| **co-occurrence+dist(F,center)** | 79.8% | 89.1% | 85.7% |
| **co-occurrence+area(F)** | 80.1% | 89.1% | 86.1% |
| **ALL** | **85.2%** | **91.2%** | **89.0%** |

Our name-face mapping model is created by identifying salient names, constraint face identification, and improved FCM clustering algorithm. To give full exhibition to the superiority of our mapping model, we have also performed a comparison between our unsupervised method and some other methods. Two approaches developed by Poignant *et al*. [11] and Bendris *et al*. [12] are analogous with ours to some extent, and we have accomplished these methods on the same dataset. We test the CNN feature and SIFT feature on every method and the CMR is also divided into three parts. The first CMR is the CMR for the faces whose names appear in the video, the second CMR is the CMR for the faces whose names do not appear in the video and the last CMR is the CMR for the whole faces. The experimental results are presented in Table 4.

**Table 4.** The comparison results between our and the other existing classic approaches.

| Approach | Face appearance descriptors | *CMR* | | |
|---|---|---|---|---|
| | | ***Faces with name*** | ***Faces without name*** | ***All faces*** |
| **Poignant *et al*.'s (2012) [11]** | **CNN** | 78.2% | 86.1% | 83.2% |
| | **SIFT** | 69.7% | 68.2% | 68.7% |
| **Bendris *et al*.'s (2014) [12]** | **CNN** | 76.3% | 88.4% | 84.0% |
| | **SIFT** | 67.4% | 70.5% | 69.5% |
| **Our Approach** | **CNN** | **85.2%** | **91.2%** | **89.0%** |
| | **SIFT** | 84.1% | 69.9% | 74.3% |

The instantiation of some name-face mapping examples are shown in the Fig. 5.

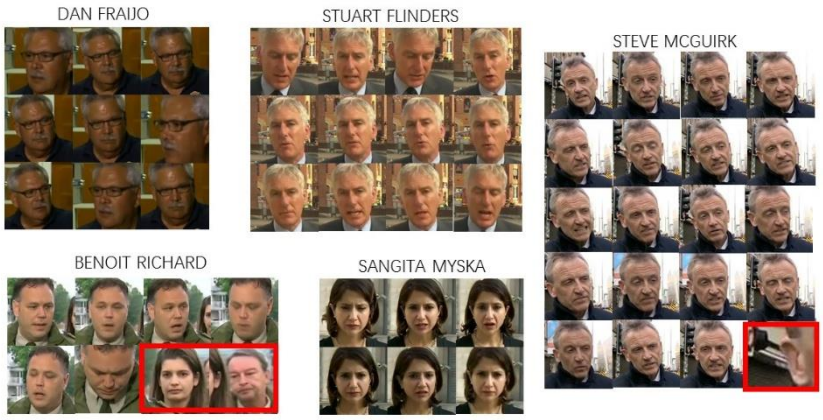


**Fig. 5.** The instantiation of some name-face mapping examples.

### 6.3 Analysis and Discussion

We can see that facial features have great influence on the result and we know that the facial features mainly affect the result of the face clustering. Although the features extracted by CNN technology is better than traditional features, we cannot obtain perfect results of the face clustering. The face clustering is difficult due to the huge variation in the appearance of each character. The extractor of the facial features is mainly trained by the static images and they are different from the faces extracted from the video. If we can train the extractor by the faces extracted from the video, we shall obtain better descriptors of the face.

It can be found from Table 4 that we can obtain the best CMR values. Poignant *et al.*'s approach tries to match all the names with their faces and Bendris *et al.*'s approach only matches the name with the speakers. Because of this, Poignant *et al.*'s CMR for the faces with names is better than Bendris *et al.*'s and Bendris *et al.*'s CMR for the faces without names is better than Poignant *et al.*'s. Our approach focuses on distinguishing the people that appear at the same time so we can obtain better CMR. Many faces often appear in the same time in videos and we can obtain better mapping results by distinguish the important person from the whole faces.

## 7 Conclusions

In this paper, a new framework is introduced to automatic annotate the person in the news video. A novel algorithm is developed by integrating identifying salient names, constraint face identification, and the improved FCM clustering algorithm. Our future

work will focus on fusing the audio information in the video and further improve the OCR accuracy.

**Acknowledgments.** This work is supported by National Natural Science Fund of China (61572140), Shanghai Municipal R&D Foundation (16511105402&16511104704), Shanghai Philosophy Social Sciences Planning Project (2014BYY009), and Zhuoxue Program of Fudan University. Yuejie Zhang is the corresponding author.

## 8 References

1. Tuytelaars T., Moens M.F. 2011. Naming people in news videos with label propagation[J]. IEEE multimedia, 18(3): 44-55.
2. Everingham M., Sivic J., Zisserman A. 2009. Taking the bite out of automated naming of characters in TV video[J]. Image and Vision Computing, 27(5): 545-559.
3. Poignant J., Besacier L., Le V.B., et al. 2013. Unsupervised naming of speakers in broadcast TV: using written names, pronounced names or both?[C]//The 14rd Annual Conference of the International Speech Communication Association, INTERSPEECH.
4. Zhou C., Zhang C., Li X., et al. 2014. Video face clustering via constrained sparse representation[C]//Multimedia and Expo (ICME), 2014 IEEE International Conference on. IEEE, 1-6.
5. Taigman Y., Yang M., Ranzato M.A, et al. 2014. Deepface: Closing the gap to human-level performance in face verification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1701-1708.
6. Parkhi O.M, Vedaldi A., Zisserman A. 2015. Deep face recognition[J]. Proceedings of the British Machine Vision, 1(3): 6.
7. Yu X., Huang J., Zhang S., et al. 2013. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model[C]//Proceedings of the IEEE International Conference on Computer Vision, 1944-1951.
8. Tuytelaars T, Moens M F. Naming people in news videos with label propagation[J]. IEEE multimedia, 2011, 18(3): 44-55.
9. Jaderberg M., Simonyan K., Vedaldi A., et al. 2016. Reading text in the wild with convolutional neural networks[J]. International Journal of Computer Vision, 116(1): 1-20.
10. Cheng Y., Liu Z., Zhao Y., et al. 2015. People news search via name-face association analysis[C]//Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, 467-470.
11. Poignant J., Bredin H., Le V.B, et al. 2012. Unsupervised speaker identification using overlaid texts in TV broadcast[C]//Interspeech 2012-Conference of the International Speech Communication Association.
12. Bendris M., Favre B., Charlet D., et al. 2014. Multiple-view constrained clustering for unsupervised face identification in TV-broadcast[C]//Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 494-498.