SPEAKER RETRIEVAL FOR TV SHOW VIDEOS BY ASSOCIATING AUDIO SPEAKER RECOGNITION RESULT TO VISUAL FACES*

Yina Han', Joseph Razik', Gerard Chollet', and Guizhong Liu**

*School of Electrical and Information Engineering, Xi'an Jiaotong University, Xi'an, China
'CNRS-LTCI, TELECOM-ParisTech, Paris, France

ABSTRACT

Person retrieval based on solely visual face recognition is hard because of the well known problems of illumination, pose, size and expression variation, which can exceed those due to identity. Fortunately, videos often accompanied with other modalities, like audio, text, etc. In this paper, we propose a framework to associate who and when information provided by speaker recognition result to the present faces in the frame sequence for retrieving speakers in TV show videos. First shot segmentation and clustering is carried out. Then face detection and tracking is followed to further locate the faces spacially. Finally, given the identity and time information by speaker recognition result, we point out three ambiguities to be resolved and propose their corresponding solutions.

1. INTRODUCTION

The goal of this work is to retrieve speakers in TV show programs using their names as the query. One approach to this problem is to use face recognition, which is a long standing and well studied problem [1]. However, as has been noted by previous authors [1], most of the face recognition methods are evaluated only in controlled environments and for relatively limited sets of faces and poses. For more realistic data sets, such as TV show programs, face recognition is extremely challenging visually as speakers exhibit significant variation in their imaged appearance due to changes in scale, pose, expressions, partial occlusion etc.

Since image/video data are often accompanied not only with images, but also with audio, captions, speech etc, it has been shown that use of multi-modality allows better retrieval and analysis recently. As in [1, 2] , combining the accompanied names from the caption and faces from the images allows better retrieval performances without requiring recognition, since the face recognition problem is simplified and transformed into the problem of finding associations between names and faces [1]. Speaker recognition from audio can give us similar cues like accompanied caption in [1, 2] by providing who and when speaks information.

*THIS WORK IS SUPPORTED BY K-SPACE AND INFOMAGIC PROJECT.

In this paper, based on the assumption that precise speaker recognition is available, we explore how to associate this result to faces present in TV show videos for speaker retrieval. Since audio speaker recognition results record who and when speaks, it appears to be considerably simpler to find associations between names and faces than it is to identify the face.

2. FRAMEWORK OVERVIEW

Our framework is mainly composed of two parts: visual processing part that focuses on clustering the same person with different occurrences together and locate their faces spacially [3]; and identity association part which focuses on resolving the ambiguities between present faces and who and when speaks information.

2.1. Shot layer processing

Based on the idea that the TV show program we focused on exhibits limited and compact view types for each main speaker, as in our previous work [3], shot segmentation and clustering is conducted first. A typical TV show program contains around hundreds of segmented shots, but these arise from just dozens of different "clusters" for different main speakers. Discovering the correspondence between shot segments reduces the volume of data to be processed, avoids the complex facial feature based inter shot person matching, and allows stronger appearance models to be built for each speaker.

2.2. Face detection

The output of the shot processing stage gives a coarse segmentation of speakers in frame layer. To further localize their faces presented in each frame, face detection is employed. Multi-view face detectors are now available. In order to give much greater reliability, an OpenCV implementation of the method of Viola and Jones [4] for frontal face detector is run on every frame, and to achieve a low false positive rate, a conservative threshold on detection confidence is used, as in [5, 6].The use of frontal face detector restricts the video content we can label to frontal faces [6], and there may be drop

(a) (b)

Fig. 1. Examples of speaker ambiguity. (a) Four faces are detected but only one person is speaking. (b) A 'reaction shot' - the speaker is not visible in the frame. (b) Also can be seen as a 'silence shot' for the person to be retrieved.

outs as the person turns towards profile and back to frontal. Hence face tracking is adopted as our solution.

2.3. Face tracking

For each shot, mean shift color tracker is applied. The mean shift algorithm has achieved considerable success in object tracking due to its simplicity and robustness. It finds local minima of a similarity measure between the color histograms or kernel density estimates of the model and target image [7]. This simple tracking procedure is extremely robust and the faces can be located with high reliability in each frame despite variation in pose, lighting, and facial expression.

2.4. Association of who and when to faces

Since a speaker is likely to appear when he/she is speaking, a certain speaker segment in audio domain can be used to limit the search space for retrieving him visually. However, we are still faced with three problems of ambiguity: (i) there might be several faces present in the frame and we do not know which one is the speaker; (ii) the actual present person might be not the speaker since this frame is just part of a 'reaction shot', (iii) the speaker to be retrieved might be present at some frames where they did not speak anything, hence there is no corresponding who and when information available. These are illustrated in Figure1. How to solve the ambiguities and retrieval all the appearances of the specific person is our next step. This includes three key technical points to be resolved:

First, how to identify the speaker from the several detected faces present in the same frame. This can be achieved by finding face detections with significant lip motion as suggested in [6].

Second, how to set reasonable association between speakers indicated in the audio speaker recognition results to the shots really contain a speaker, namely to the shot which is not a reaction one. Since a speaker shot is likely to frequently appear when he/she is speaking, the duration information for each shot within a specific speaker segment is an important cue.

Third, how to set shot layer visual representation for each speaker so as to retrieve his/her 'silent' appearance. Visual similarity matching between determined faces and undetermined ones based on either global face features, like eigenface, or local facial feature exemplars, like SIFT [5, 6], need furhter study.

3. CONCLUSION

In this paper we propose a framework for speaker retrieval based on both audio speaker recognition results and visual information. Our goal is more restricted than general face recognition in that we need only set association for faces present in the video to their names provided by audio speaker recognition results. Hence it is a method to increase the retrieval performance of person queries in TV show videos where a precise audio speaker recognition result was provided in advance and where traditional face recognition systems cannot be used. It does not require a training step for a specific person and therefore, there is no limit on the number of people queried.

4. REFERENCES

[1] D. Ozkan and P. Duyqulu, "A graph based approach for naming faces in news photos," in *Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1477–1482.

[2] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Teh Yee-Whye, E. Learned-Miler, and D.A. Forsyth, "Names and faces in the news," in *Computer Vision and Pattern Recognition (CVPR)*, 2004, pp. 848–854.

[3] Y. Han, G. Liu, G. Chollet, and J. Razik, "Person identity clustering in tv show videos," in *Accepted by Visual Information Engineering (VIE)*, 2008.

[4] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 52, no. 2, pp. 137–154, 2004.

[5] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: video shot retrieval for face sets," in *International Conference on Image and Video Retrieval (CIVR)*, 2005, pp. 226–236.

[6] M. Everingham, J. Sivic, and A. Zisserman, ""hello! my name isbuffy"- automatic naming of characters in tv video," in *Proceedings of British Machine Vision Conference (BMVC)*, 2006.

[7] C. Yang, R. Duraiswami, and L. Davis, "Efficient mean-shift tracking via a new similarity measure," in *Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 176–183.