HCMUS team at the Multimodal Person Discovery in Broadcast TV Task of MediaEval 2016

Vinh-Tiep Nguyen, Manh-Tien H. Nguyen, Quoc-Huu Che, Van-Tu Ninh,
Tu-Khiem Le, Thanh-An Nguyen, Minh-Triet Tran
Faculty of Information Technology
University of Science, Vietnam National University-Ho Chi Minh city
nvtiep@fit.hcmus.edu.vn, {nhmtien, cqhuu, nvtu, ltkhiem}@apcs.vn,
1312016@student.hcmus.edu.vn, tmtriet@fit.hcmus.edu.vn

ABSTRACT

We present the method of the HCMUS team participating in Multimodal Person Discovery in Broadcast TV Task at the MediaEval Challenge 2016. There are two main processes in our method. First we identify a list of potential characters of interest from all video clips. Each potential character is defined as a pair of face track, a sequence of face patches, and a name. We use OCR results and face detection to find potential characters. We also apply several simple techniques to check the consistency of linking a name with a face track to reduce potential wrong matching pairs. Then we detect face patches from test video shots with cascade DPM, extract deep features from face patches using a very deep Convolutional Neural Network, and classify faces using SVM.

1. INTRODUCTION

The objective of the Multimodal Person Discovery in Broadcast TV Task is to automatically find the appearance of main characters from a large dataset of broadcast TV clips [1]. Name of a person can be introduced by text in caption, or via speech in conversation.

In our approach to solve this task, we use two types of data: text from caption and visual data. There are two main processes in our method: (i) identify characters of interest and their names; and (ii) discover person using face recognition with deep features.

In the first process, we propose the Main Text Verification step based on font size to select text phrases that may be used with high confidence as character names. If there are multiple main phrases detected in a single frame, we consider it as an ambiguous frame and eliminate it. Then we extract faces at the timestamps corresponding to a found name with OCR. Only faces with significant size are chosen to link with names. Besides, a frame containing multiple faces with large size are discarded as we may not associate a face correctly with a name. Finally, if a group of faces associated with a single name has different large sub-groups of faces of multiple persons, this group is not reliable and should be discarded.

After the first process, we obtain a list of evidence entries containing names and associated faces. In the second process, we use VGG-Face [2] to extract deep features from face patches and train SVM models with these 4096-dimension features to classify all main characters found in the first process. We use Cascade DPM [3,4] to detect and crop facial areas from frames to boost the accuracy of SVM classifiers. For each face patch extracted from a test video shot, we classify it with the trained SVM models.

2. IDENTIFY CHARACTERS OF INTEREST

There are many persons in video clips. However, we only focus on main characters – person appearing clearly on a frame with names explicitly introduced either by caption text or speech.

The process to identify characters of interest includes three main phases: detect names, identify <name, face> pairs, and verify consistency of <name, face> pairs.

In name detection phase (c.f. Figure 1), caption text is first extracted from video frames with OCR module. Main Text Verification step is used to filter text phrases that may not be used with high confidence as character names. We apply simple techniques for name detection and finally get a list of names and corresponding timestamps (in video shots).

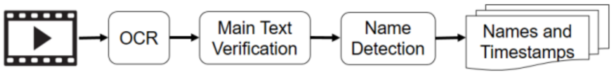
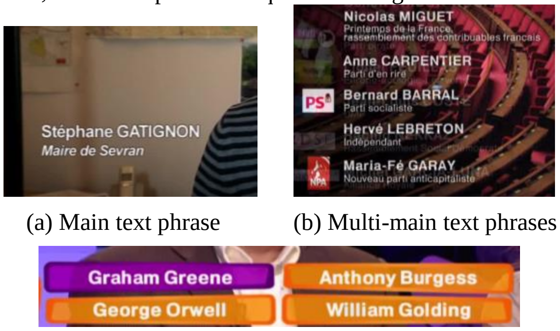


Figure 1. Name Detection Process

In Main Text Verification module, as the name of a character is usually displayed with a large font, we eliminate text phrases with small size. Only text phrases with the largest font in a frame are selected (Figure 2a). Besides, we also discard frames having multiple text phrases with the same largest size because we may not link a name with a detected main face in such frames with high confidence (Figure 2b and c). This situation often occurs in the introduction of a show, in a scrolling list of person at the end of a film, or in multiple-choice question of a gameshow.



(a) Main text phrase (b) Multi-main text phrases

(c) Names in a question of gameshow

Figure 2. Main text verification

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

Face patches are extracted from each frame corresponding to the timestamp of a potential name. As the name of a character usually appears when the face of that person can be seen clearly in frontal pose, we use Viola-Jones face detector[5] in this step. In the Main Face Verification step, a face is considered as a main face if its size is large enough. In our experiment, we choose the threshold value for a main face as 7% of the total frame's area. In this step, a frame with more than one main face is discarded as we may not link correctly a name with a main face (Figure 5a). The output of this step is a list of <Name, Face> pairs.

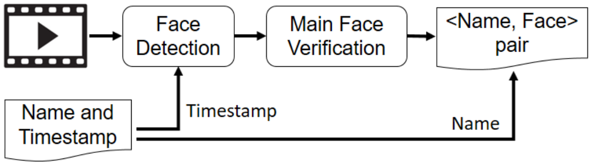


Figure 3. <Name, Face> Pair Identification Process

In Face Consistency Verification process (Figure 4), if there are many different faces associated with a given name, we discard such <Name, Face> pairs. An example of such situation is that the name is of a program, not a person and the name appears throughout the program (Figure 5b).

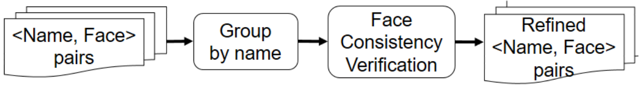
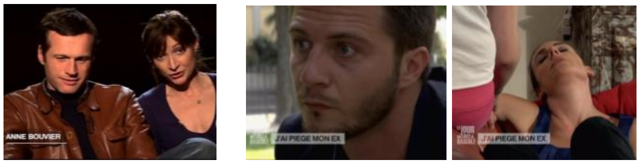


Figure 4. Face Consistency Verification



(a) Multiple main faces (b) Program's Name and Multiple Person

Figure 5. Difficult Cases for <Name,Face> Pairs

3. FACE RECOGNITION USING DEEP FEATURES

We propose to use face recognition to find shots containing person whose names were recognized via caption text using OCR algorithm. Starting from an evidence entry of a video that contains only one person (to make sure that the face and name are associated), we keep track face bounding boxes of next video frames. For this face track, we extract deep features from face patches using a very deep Convolutional Neural Network (VGG-Face [2])

After this module, each face patch will be represented by a 4096-dimensional feature vector. Although this feature is designed to best fit with L_2 distance metric, there still is a big gap in performance. This could be explained by the fact that the face feature vector does not have the same weight for all components. For each face, the weights of components are different. Therefore, we propose to learn these features by a large margin classifier. All features of an evidence face are collected for training a Support Vector Machine (SVM) algorithm with linear kernel.

For negative examples, we collect face features from other persons in the evidence file. To further improve the recognition performance, we use cross-validation method with k=5 folds. After this step, each person of an evidence entry will be represented by a face classifier. This classifier is used to recognize person appearing in all test video shots counted from current face track. We apply Cascade DPM detector [3,4] on image region near the face track of metadata to extract face patches which then be transferred to VGG-Face network. Using this detector instead of other face detectors, such as Viola-Jones one, improves the performance because the network is trained on face patches extracted by this algorithm.

In testing phase, each face candidate detected in a frame of a shot is transferred to the face classifier and scored by a confidence value. A positive value means candidate face is likely to be similar to the person of trained on classifier and vice versa.

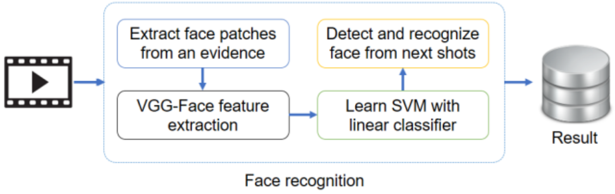

Extract face patches from an evidence
Detect and recognize face from next shots
VGG-Face feature extraction
Learn SVM with linear classifier
Result
Face recognition

Figure 6. Person Discovery with Face Recognition

4. CONCLUSION AND FUTURE WORK

In our current approach, we focus our effort in refining names of main characters appearing in broadcast TV clips, and applying VGG-Face and SVM to recognize face patches extracted with cascade DPM. As audio data is not utilized in our method, we miss information about a person from speech in conversation. Thus, we will continue to exploit this data type to obtain potential information about person introduction.

Besides, in our method, as we want to boost the accuracy of the trained SVM classifiers, we use cascade DPM to extract face patches. This process, together with extracting deep features with VGG-Face, is time consuming. Therefore, by the end of the challenge, we still have shots to be processed. Currently we are revising our method and will use it to process again the whole data from this challenge.

5. ACKNOWLEDGEMENT

We would like to express our appreciation to Multimedia Processing Lab, University of Information Technology, VNU-HCM, and Computational Imaging Group at University of Illinois at Urbana-Champaign for their supports of computing infrastructure to our team in this challenge.

6. REFERENCES

[1] Bredin, H., Barras, C., Guinaudeau, C. 2016. Multimodal Person Discovery in Broadcast TV at MediaEval 2016. In *Proc. of the MediaEval 2016 Workshop*, Hilversum, Netherlands, Oct. 20-21, 2016.

[2] Parkhi, O. M., Vedaldi, A., Zisserman, A. 2015. Deep Face Recognition. In *Prof. of British Machine Vision Conference (BMVC) 2015*

[3] Wolf, L., Hassner, T., Maoz, I. 2011. Face Recognition in Unconstrained Videos with Matched Background Similarity. 2011. In *Proc. of Computer Vision and Pattern Recognition (CVPR) 2011*.

[4] Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L. 2014. Face detection without bells and whistles. In *Proc. of European Conference on Computer Vision (ECCV) 2014.*

[5] Viola, P., Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Proc. of Computer Vision and Pattern Recognition* (CVPR) 2001.