GTM-UVigo System for Multimodal Person Discovery in Broadcast TV Task at MediaEval 2016

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo
Multimedia Technologies Group (GTM), AtlantTIC Research Center, University of Vigo
E.E. Telecomunicación, Campus Universitario S/N, 36310 Vigo
{plopez,ldocio,carmen}@gts.uvigo.es

ABSTRACT

In this paper, we present the system developed by GTM-UVigo team for the Multimedia Person Discovery in Broadcast TV task at MediaEval 2016. The proposed approach consists in a novel strategy for person discovery which is not based on speaker and face diarisation as in previous works. In this system, the task is approached as a person recognition problem: there is an enrolment stage, where the voice and face of each discovered person are detected and, for each shot, the most suitable voice and face are assigned using the i-vector paradigm. These two biometric modalities are combined by decision fusion.

1. INTRODUCTION

The Person Discovery in Broadcast TV task at MediaEval 2016 aims at finding out the names of people who can be both seen as well as heard in every shot of a collection of videos [2]. This paper describes a novel approach that is not based on speaker and face diarisation as is usually done in this task [6, 7, 8, 10]; instead, the task is approached as a person recognition problem.

2. SYSTEM DESCRIPTION

The proposed system can be divided in an enrolment and a search stage. For each person name detected by optical character recognition (OCR), the most likely interval of speech and face presence are detected and used for enrolment. Once the detected people are enrolled, speaker and face recognition are performed for each shot in order to assign labels to that shot. A decision fusion strategy is implemented in order to combine the speech and video labels. The details of the system are described below.

2.1 Name detection

The person names were obtained from the video using the baseline system provided by the organisers. Specifically, the UPC OCR approach using LOOV was used [11]. Since the output of the OCR module had errors such as including additional words in the person name, a naïve filtering of the OCR output was performed by removing those names that had more than four words.

2.2 Speech enrolment

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

First, features were extracted from the waveform; specifically, 19 Mel-frequency cepstral coefficients (MFCCs) including energy were extracted every 10 ms using a 25 ms sliding window. A dynamic normalisation of the cepstral mean was applied using a sliding window of 300 ms. These features were extracted using the Kaldi toolkit [12]. Then, for each person name detected by the OCR:

- The time interval $(t_{\text{start}}, t_{\text{end}})$ in which the name of the speaker *spk* appears is taken as a starting point. A strategy to enlarge this time interval in order to obtain more data to enrol the speaker is applied: given the time intervals $S_{\text{left}} = (t_{\text{start}} - 10, t_{\text{end}})$ and $S_{\text{right}} = (t_{\text{start}}, t_{\text{end}} + 10)$, a change point is searched within each of these intervals using the Bayesian information criterion algorithm (BIC) for speaker segmentation, having the restriction that the change point has to be in the intervals $(t_{\text{start}} - 10, t_{\text{start}})$ and $(t_{\text{end}}, t_{\text{end}} + 10)$, respectively. If no change point was found within the interval S_{left} then t_{left} is set to $t_{\text{start}} - 10$ and, similarly, if no change point was found within the interval S_{right} then t_{right} is set to $t_{\text{end}} + 10$. Then, speaker *spk* is assumed to be speaking in the interval $S_{\text{spk}} = (t_{\text{left}}, t_{\text{right}})$. In case speaker *spk* appears several times in the OCR output, a segment is computed for each occurrence.

- Speech activity detection (SAD) was performed in order to remove the non-speech parts. To do so, the energy-based SAD approach implemented in the Kaldi toolkit was applied.

- An i-vector [5] was extracted for speaker *spk* using the Kaldi toolkit. In case several segments were obtained in the first step, their features were concatenated and all the segments were treated as a single one. In this step, the 19 MFCCs were augmented with their delta and acceleration coefficients.

2.3 Face enrolment

When dealing with faces, the first step consisted in performing face tracking using the baseline approach based on histogram of oriented gradients [3] and the correlation tracker proposed in [4]. Then, for each person name detected by the OCR:

- The faces detected by the face tracker in the interval $(t_{\text{start}}, t_{\text{end}})$ in which the name of the speaker *spk* appears are considered. Given that only one face was detected, the whole presence interval of that face is taken. In case more than one face was detected, the

Table 1: Results achieved on the whole test data and on each partition.

	All			3-24			DW			INA		
	MAP@1	MAP@10	MAP@100	MAP@1	MAP@10	MAP@100	MAP@1	MAP@10	MAP@100	MAP@1	MAP@10	MAP@100
p	0.315	0.236	0.211	0.538	0.394	0.366	0.242	0.185	0.185	0.358	0.265	0.208
c1	0.293	0.182	0.168	0.487	0.338	0.314	0.242	0.157	0.157	0.314	0.178	0.146
c2	0.245	0.199	0.177	0.333	0.303	0.286	0.116	0.088	0.088	0.302	0.170	0.132
b	0.363	0.273	0.247	0.667	0.477	0.462	0.251	0.186	0.186	0.440	0.341	0.276

one that appeared in more frames was assigned to the speaker, assuming that was the dominant face in the given time interval.

- Features were extracted in the time interval obtained in the previous step. To do so, first face detection was performed, and a geometric normalisation was done. After that, photometric enhancement of the image using the Tan&Triggs algorithm [13] was applied. Finally, discrete cosine transform features (DCT) [9] were extracted using blocks of size 12 with 50% overlap and 45 DCT components. The feature extraction stage was performed using the Bob toolkit [1].

- Once the features were obtained, an i-vector representing that face was obtained using the Kaldi toolkit. As done when dealing with speech, if there were several time intervals where the face of the speaker was present, the features obtained in all the segments were concatenated.

2.4 Search

The procedure to decide which speaker was present in each shot consisted in, for each shot:

- In order to detect whether the shot includes speech, speech detection was performed: perceptual linear prediction coefficients plus pitch features were extracted from the time interval defined by the shot, an i-vector was extracted and a logistic regression approach was used to classify the segment as speech or non-speech. Non-speech segments were straightforwardly discarded. In case speech was present in the shot, SAD was performed, an i-vector was extracted, and this shot i-vector was compared with the enrolment i-vectors computing the dot scoring. The speaker that achieved the highest score was assigned to the shot.

- The faces detected by the face tracker within the shot were identified, and the one that appeared in more frames was chosen as the most representative face of the shot. An i-vector was extracted and the same decision procedure described for the speech data was performed.

- Once a decision was made for both speech and face data, the following fusion approach was implemented: given a shot, it is assigned to a speaker if the person detected by the face and speech detectors had the same name and if the sum of their scores was greater than a threshold.

3. RESULTS AND DISCUSSION

Table 1 shows the results achieved with the audio+video fusion system (p), the audio system only (c1), the video system only (c2) and the baseline provided by the organisers (b). The main conclusions that can be extracted from the Table are: (1) the audio and video systems are complementary, since their combination leads to an improvement of the individual results; (2) the audio results are better than the video results, especially in the DW database; and (3) the worst results were obtained in the DW database, while the best ones were achieved in the 3-24 database. The reason why 3-24 results are, in general, better, might be caused by the small number of queries in the evaluation data corresponding to this database (only 15 queries out of 693), which leads to results that are not significative. In the case of DW database, 606 queries were evaluated; this, combined with the fact that the OCR approach used in this system did not find person names in 612 out of 757 files in the database, led to poor results in DW data.

The aim of this system was to assess a novel approach for person discovery that is not based on speaker and face diarisation as in most state-of-art strategies. The achieved results are promising, and the experiments performed in this evaluation allowed the detection of the main weak points of the system that will be improved in the future:

- The quality of the OCR output had a huge impact on the results, since this is the starting point of the whole enrolment stage, which leads to a degradation of performance on the whole system. A simple approach, based on natural language processing, for filtering the OCR output in order to remove everything that were not person names was assessed in this framework with no success, but further experiments on this topic will be done in the future.

- All face-based steps relied on the baseline approach for face tracking, and its output was fed to the feature extraction module; however, only the information about presence was used, but not the bounding boxes where the faces appeared. This probably led to inconsistencies in the feature extraction stage and, therefore, on the face enrolment procedure. This issue will be addressed in order to improve the quality of the face-based approach.

Acknowledgements.

This research was funded by the Spanish Government under the project TEC2015-65345-P, the Galician Government through the research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014) and 'AtlantTIC Project' CN2012/160, and by the European Regional Development Fund (ERDF).

4. REFERENCES

[1] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *20th ACM Conference on Multimedia Systems (ACMMM)*, 2012.

[2] H. Bredin, C. Barras, and C. Guinaudeau. Multimodal Person Discovery in Broadcast TV at MediaEval 2016. In *Proceedings of the MediaEval 2016 Workshop*, 2016.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, 2005.

[4] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2014.

[5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[6] M. India, D. Varas, V. Vilaplana, J. Morros, and J. Hernando. UPC system for the 2015 MediaEval multimodal person discovery in broadcast TV task. In *Proceedings of the MediaEval 2015 Workshop*, 2015.

[7] N. Le, D. Wu, S. Meignier, and J.-M. Odobez. EUMSSI team at the MediaEval person discovery challenge. In *Proceedings of the MediaEval 2015 Workshop*, 2015.

[8] P. Lopez-Otero, R. Barros, L. Docio-Fernandez, E. Gonzalez-Agulla, J. Alba-Castro, and C. Garcia-Mateo. GTM-UVigo systems for person discovery task at MediaEval 2015. In *Proceedings of the MediaEval 2015 Workshop*, 2015.

[9] C. McCool and S. Marcel. Parts-based face verification using local frequency bands. In *Proceedings of IEEE/IAPR international conference on biometrics*, 2009.

[10] F. Nishi, N. Inoue, and K. Shinoda. Combining audio features and visual i-vector @ MediaEval 2015 multimodal person discovery in broadcast TV. In *Proceedings of the MediaEval 2015 Workshop*, 2015.

[11] J. Poignant, L. Besacier, G. Quénot, and F. Thollard. From text detection in videos to person identification. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2012.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE Signal Processing Society, 2011.

[13] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650, 2010.