UPC System for the 2016 MediaEval Multimodal Person Discovery in Broadcast TV task*

Miquel India, Gerad Martí, Carla Cortillas, Giorgos Bouritsas
Elisa Sayrol, Josep Ramon Morros, Javier Hernando
Universistat Politècnica de Catalunya

ABSTRACT

The UPC system works by extracting monomodal signal segments (face tracks, speech segments) that overlap with the person names overlaid in the video signal. These segments are assigned directly with the name of the person and used as a reference to compare against the non-overlapping (unassigned) signal segments. This process is performed independently both on the speech and video signals. A simple fusion scheme is used to combine both monomodal annotations into a single one.

1. INTRODUCTION

This paper describes the UPC system for the 2016 Multimodal Person Discovery in Broadcast TV task [2] in the 2016 MediaEval evaluations. The system detects face tracks (FT), detects speech segments (SS) and also detects the person names overlaid in the video signal. Both the video and the speech signals are processed independently. For each modality, we aim to construct a classifier that can determine if a given FT or SS belongs or not to one of the persons appearing on the scene with an assigned overlaid name. As the system is unsupervised, we will use the detected person names to identify the persons appearing on the program. Thus, we assume that the FT of SS that overlap with a detected person name are true representations of this person.

The signal intervals that overlap with an overlaid person name are extracted and used for unsupervised enrollment, defining a model for each detected name. This way, a set of classes corresponding to the different persons in the detected names is defined. These intervals are directly *labeled* by assigning the identity corresponding to the overlaid name.

For each modality, a joint identification verification algorithm is used to assign each *unlabeled* signal interval (FT or SS not overlapping with an overlaid name) to one of the previous classes. For each unlabeled interval, the signal is compared against all models and the one with better likelihood is selected. An additional 'Unknown' class is implicitly considered, corresponding to the cases where the face track or speech segment correspond to a person that is never named (i.e. none of the appearances of this person in the video do overlap with a detected name).

At the end of this process we have two different sets of annotations, one for speech and one for video. The two results are fused, as described in section 5 to obtain the final annotation.

2. TEXT DETECTION

We have used the two baseline detections with some additional post-processing. The first one (*TB1*) was generated by our team and distributed to all participants. The LOOV [6] text detection tool was used to detect and track (define the temporal intervals where a given text appears) text. Detections were filtered by comparing against list of *first names* and *last names* downloaded from the internet. We also used lists of *neutral particles* ('van', 'von', 'del', etc.) and *negative names* ('boulevard', etc.). All names were normalized to contain only alphabetic ASCII characters, without accents nor special characters and in lower case. For a given detected text to be considered as name it had to contain at least one first name and one last name. The percentage of positive matches for these two classes was used as a score. Matches from the neutral class did not penalize the percentage. Additionally, if the first word in the detected text was included in the negative list, the text was discarded. To construct *TB1* we had access to the test videos before than the rest of participants. However, we only used this data for this purpose and we did not perform any test of the rest of our system before the official release.

The second set of annotations, *TB2* was provided by the organizers [2]. These annotations had a large quantity of false positives. We applied the above described filtering to *TB2* and we combined the result with *TB1*, as the detectors resulted to be partly complementary.

3. VIDEO SYSTEM

For face tracking, the 2015 baseline code [7] was used. Filtering was applied to remove tracks shorter than a fixed time or with too small face size.

The VGG-face [8] Convolutional Neural Network (CNN) was used for feature extraction. We extracted the features from the activation of the last fully connected layer. The

*This work has been developed in the framework of the projects TEC2013-43935-R, TEC2012-38939-C03-02 and PCIN- 2013-067. It has been financed by the Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund (ERDF).

MediaEval 2016 workshop Octobber 19–21, 2016, Hilversum, Netherlands

ACM ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235

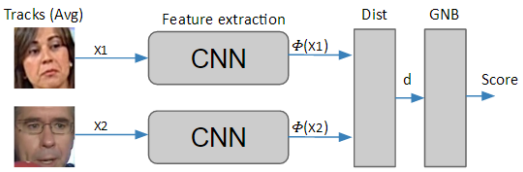

Figure 1: Diagram of the verification system

network was trained using a triplet network arquitecture [5]. The features from the detected faces in each track are extracted using this network, and then averaged to obtain a feature vector for each track, of size 1024.

A face verification algorithm was used to compare and classify the tracks. First, the tracks that were overlapped with a detected name were named by assigning that identity. To reduce wrong assignations, the name was only assigned if it overlapped with a single track. Then, using the set of named tracks from the full video corpus, a Gaussian Naive Bayes (GNB) binary classifier model was trained, using the euclidean distance between pairs of samples from the named tracks. Then, for each specific video, each unnamed track was compared with all the named tracks of the video, computing the euclidean distance between the respective feature vectors of the tracks (see Figure 1). This value was classified using the GNB to either being a intra-class distance (both tracks belong to the same identity) or an inter-class distance (the tracks are not from the same person). The probability of the distance being intra-class was used as the confidence score. The unnamed track was assigned the identity of the most similar named track. A threshold on the confidence score (0.75) was used to discard tracks not corresponding to any named track.

4. SPEAKER TRACKING

Speaker information was extracted using an i-vector based speaker tracking system. Assuming that text names are temporarily overlapped with their speaker and face identities, speaker models were created using the data of those text tracks. Speaker tracking was performed evaluating the cosine distance between model i-vectors and i-vectors extracted for each frame of the signal.

Speaker modelling was implemented using i-vectors [3]. An i-vector is a low rank vector, typically between 400 and 600, representing a speech utterance. Feature vectors of the speech signal are modeled by a set of Gaussian Mixtures (GMM) adapted from a Universal Background Model (UBM). The mean vectors of the adapted GMM are stacked to build the M supervector, wich can be written as:

$$M = m_u + T\omega \tag{1}$$

where m_u is the speaker- and session-independent mean supervector from UBM, T is the total variability matrix, and ω is a hidden variable. The mean of the posterior distribution of ω is referred to as i-vector. This posterior distribution is conditioned on the Baum-Welch statistics of the given speech utterance. The T matrix is trained using the Expectation-Maximization (EM) algorithm given the centralized Baum-Welch statistics from background speech utterances. More details can be found in [3].

The speaker tracking system has been implemented as a speaker identification system with a segmentation by classification method. For the feature extraction, 20 Mel Frequency Cepstral Coefficients (MFCC) plus Δ and $\Delta\Delta$ coefficients were extracted. Using the Alize toolkit[4, 1], a total variability matrix has been trained per show. I-vectors have been extracted from 3 seconds segments with a 0.5 second shift and the baseline speaker diarization was used to select speaker turn segments to extract the i-vector queries. The identification was performed evaluating the cosine distance of the i-vectors with each query i-vector. The query with the lowest distance was assigned to the segment. A global distance threshold was previously trained with the development database so as to discard assignations with high distances.

System	MAP1(%)	MAP5(%)	MAP10(%)
Baseline 1	13.1	12	11.7
Spk Tracking	43.3	30.6	28.8
Baseline 2	37	30.3	29.2
Face Tracking	61.3	47.9	45.5
Baseline 3	36.3	29.3	27.3
Intersection	47.9	34	32
Union	**63.0**	**50.5**	**48.4**

Table 1: MAP Evaluation

5. FUSION SYSTEM AND RESULTS

Starting off with the speaker and face tracking shot labeling, two fusion methods were implemented. The first method was the intersection of the shots of both tracking systems, averaging the confidence of the intersected shots. In the second method, the union strategy was implemented relying on the intersected shots of both modalities and reducing the confidence of those not intersected. The shots of both video and speaker systems were merged, averaging the confidence score if both systems detect the same identity in a shot, or reducing the confidence by a 0.5 factor if only one of the systems detected a query.

Four different experiments were performed which are shown in Table 1. Baseline 1 refers to the fusion between the baseline speaker diarization and OCR, Baseline 2 refers to the fusion between the face detection and the OCR and Baseline 3 is the intersection of the both previous baselines. Initially, speaker and face tracking have been evaluated separately. The intersection and the union of both tracking systems were implemented as fusion strategies.

As is shown in Table 1, both monomodal systems improve the baseline performances by a great margin. The union strategy has shown a better performance than the intersection strategy but this fusion does not show a significative performance increase over the individual modalities.

By analysing the results, we believe that failures at text detection was the main factor impacting the final scores.

6. CONCLUSIONS

Speaker and face tracking have been combined using different fusion strategies. This year, our idea was to focus only on the overlaid names to develop tracking systems instead of performing diarization systems merged with text dectection. Tracking systems have shown a better performance than the diarization based ones of the baseline. For fusion, the union strategy has shown higher results than the intersection method.

7. REFERENCES

[1] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason. ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. In *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2008.

[2] H. Bredin, C. Barras, and C. Guinaudeau. Multimodal person discovery in broadcast tv at mediaeval 2016. In *Working Notes Proceedings of the MediaEval 2016 Workshop*, 2016.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, May 2011.

[4] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, H. L. Christophe Lévy, J. S.D, Mason, and J.-Y. Parfait. ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition. In *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013.

[5] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[6] J. Poignant, L. Besacier, G. Quénot, and F. Thollard. From text detection in videos to person identification. In *ICME 2012*, 2012.

[7] J. Poignant, H. Bredin, and C. Barras. Multimodal person discovery in broadcast tv at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.