UPC System for the 2015 MediaEval Multimodal Person Discovery in Broadcast TV task

M. India, D. Varas, V. Vilaplana, J.R. Morros, J. Hernando
Universitat Politecnica de Catalunya, Spain

ABSTRACT

This paper describes a system to identify people in broadcast TV shows in a purely unsupervised manner. The system outputs the identity of people that appear, talk and can be identified by using information appearing in the show (in our case, text with person names). Three types of monomodal technologies are used: speech diarization, video diarization and text detection / named entity recognition. These technologies are combined using a linear programming approach where some restrictions are imposed.

1. INTRODUCTION

The 2015 Multimodal Person Discovery in Broadcast TV [13] goal is to identify people appearing and speaking in TV shows in a purely unsupervised manner. This paper describes the UPC contribution, which is based on combining speech diarization, video-based face diarization and text detection plus Named Entity Recognition (NER). We did not make use of the names present in speech transcriptions.

2. AUDIO SYSTEM

Speaker information was extracted using an Agglomerative Hierarchical Clustering diarization system based in Hidden Markov Models [21, 20, 2, 11]. It uses energy-based speech activity detection , Mel Frequency Cepstral Coefficients voice features and initial uniform segmentation.

Speaker clusters are modeled with Gaussian Mixture Models (GMM). The complexity selection of the models is based on the amount of data per cluster and the cluster complexity ratio which fixes the amount of speech per Gaussian. Hidden Markov Model (HMM) training and cluster realignment by Viterbi decoding is based on maximum likelihood. In the decoding stage, a minimum speaker segment duration of 3 seconds is imposed to deal with too short segments. For the cluster merging, the most likely pair of clusters are selected in each iteration. This likelihood is calculated using a modified Bayesian information criterion (BIC) [4, 1] metric among clusters.

This system has been used with two different kind of inputs for each show. In one hand, diarization is run with each audio file without any constraint. In the other hand, using a face-tracking system, segments without tracked faces are discarded. The purpose of this second method is to run the diarization only in those parts where we assume that someone in the video must be speaking.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

3. VIDEO SYSTEM

For face tracking, the baseline code was used (tracking by detection using the Kanade-Lucas-Tomasi algorithm [18, 10, 16]). For feature extraction we used the technique in the baseline (HOG [5] features on facial locations[19], concatenated and projected using LDML [8]). While in the baseline a single descriptor was selected for each track, we used several vectors, by uniform temporal sampling of the track faces. We expect this approach to better capture the variations in pose/expression.

We used agglomerative hierarchical clustering. A binary hierarchical tree is created by fusing tracks according to the minimum distance between track vectors. The number of clusters may vary between videos and has to be determined. It is estimated by evaluating the CalinskiHarabasz [3] and Silhouette [14] criteria in the range $[50, 80]$ clusters and averaging the maximum results. The number of resulting clusters is the average of the maximum result for both methods.

To improve the diarization, spatio-temporal restrictions were introduced. We assume that a person can not appear twice in a frame so tracks with temporal overlapping should represent different persons and are prevented to merge into the same cluster. Also, as we use a multi-vector representation for each track, vectors in the same track must be part of the same cluster. Restrictions are modeled using a matrix expressing the relationship between the feature vectors. Entries for vectors in different tracks were assigned a value of 1, entries for vectors in the same track were assigned a value $0 < v \ll 1$, and entries for vectors on temporally co-occurring tracks received a very large value $v \gg 1$. This matrix is used to point-wise multiply the vector-to-vector distance matrix used for clustering.

4. TEXT SYSTEM

We used the person names provided in the baseline [6, 12] and our own technology for obtaining person names (in different runs). From the input image a segmentation is created with a Binary Partition Tree [15] using color and stroke width [7]. A partition is built were each character is a connected component while background regions are merged. Next, regions are filtered by a sequence of binary classifiers that reject non-character components. Components accepted by the classifiers as character candidates are combined into pairs and pairs are combined into chains. A post-processing stage is applied to find missing compo-

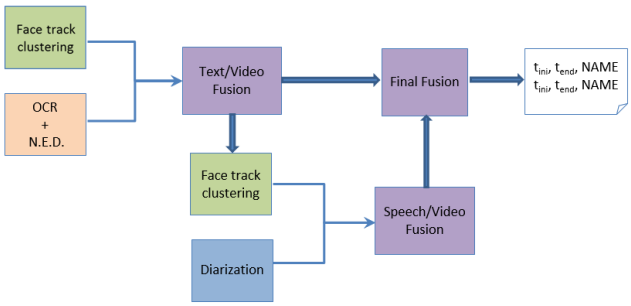

Figure 1: System block diagram

nents wrongly rejected as false positives in the filtering stage. Tesseract OCR Engine [17] provides one transcription for each text chain and Stanford Name Entity Recognizer [9] is used to automatically detect person names in the text.

5. FUSION

Our system combines the previous information sources to obtain the final person recognition labelling. Speaker diarization and video diarization are performed first in an independent manner. In order to fuse this information to create a final labelling, the development database was analyzed. Some assumptions were made:

- Speaker is not always related with who is shown in the screen. So it is important to weigh accurately the temporal overlaps between each speaker with its different possible face identity assignments.
- Some speakers do not come into view any time in the show and there are other people who are shown in the screen but do not speak. Both should be discarded.
- Text identities are more related with who is shown rather than with who is speaking. So text is better combined with video than with speech.

According to these assumptions, an algorithm was designed based in weighting temporal overlaps between tracks (Figure 1). This algorithm considers two different fusion modalities (Video/Text and Video/Audio) and combines both to obtain a final track file. Firstly, text and video are fused. Their overlapped tracks are selected, and the temporary overlaps of their identities are weigthed to set the constraints of an ILP system (IBM CPLEX).

$$\max_{\alpha_{ij}}(\sum_{i}\sum_{j}\alpha_{ij}\beta_{ij}) \quad (1)$$

$$\sum_{j}\alpha ij \leq 1 \quad (2)$$

(α_{ij}: assignment between i text identity with a j video identity; β_{ij}: weight of assignment). Equation 2 establishes that each text identity must only have one face identity assigned. The next step is to combine the speech diarization tracks with the face tracks that have a text identity assigned. The same method based on ILP is used. Finally, using the relation between text, face and speaker identities and the overlapped tracks in the second fusion, the final labeling output was obtained. A second algorithm was implemented changing the order of the fusions. In this case, audio were fused with the video and the result was combined with the text identities. Thus, only the face identities with a speaker assigned were considered.

Exp.	System	Audio Input	NER	MAP
1	2	facetrack	Baseline	22.6
2	1	facetrack	Baseline	27.1
3	2	-	Baseline	33.5
4	**1**	**-**	**Baseline**	**41.6**
5	1	-	UPC system	32.6

Table 1: MAP Evaluation

6. RESULTS

Five different experiments were performed, which are shown in Table 1. These experiments were evaluated with the training database and evaluated using the mean average precision metric (MAP). In the experiments we tested several variations: the order of the fusions, the input of the audio diarization and the text system used. In Table 1, System 1 refers to the architecture shown in Figure 1 where the first fusion combines text and video, and System 2 refers to first combining video and audio and later fusing text. *facetrack* indicates that the audio diarization is performed using only audio tracks where there are faces detected. The null case means performing the diarization using the whole audio input. While the first four experiments use the baseline names, in the fifth one the system described in section 4 was used.

The best performance was achieved in experiment 4 by the System 1, without filtering the audio input for the diarization and using the Baseline person names. There is a clear evidence that the system works better when the diarization is run with the whole audio input. Referring to the fusion order in the algorithm, results indicate that mixing video and text tracks first, provides a better performance.

The five experiments were run on the test data. Experiments 1-4 were submitted on July 1st and experiment 5 on July 8th. The best set-up in the training data (Exp.4 in Table 1) was uploaded as our primary submission. After evaluating this primary submission with the final set of annotations, the following results were obtained: EwMAP = 54.1%, MAP = 54.36% and C = 69.71%. Experiment 5 was submitted on July 8th. It is similar to experiment 4 but using our own technology to obtain person names. We had low performance with the OCR and NER and thus the results were worse than expected.

7. CONCLUSIONS

Speaker diarization, face recognition, and text detection with named entity recogniton have been combined using the integer linear programming approach. Our idea was to first perform monomodal speech and video diarizations, using as much restrictions as possible to improve the results and then use ILP to combine these diarizations along with the persons name information. Several architectures for this combination and several constrains of the integer linear programming algorithm were considered. The architecture which combines video and audio modalities after the fusion with the text stream has provided the best results.

8. ACKNOWLEDGMENTS

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).

9. REFERENCES

[1] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. *Proc. ASRU*, 2003.

[2] X. Anguera, C. Wooters, and J. Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022, 2007.

[3] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1):1–27, 1974.

[4] S. S. Chen and P. Gopalakrishnan. Clustering via the bayesian information criterion with applications in speech recognition. *Proc. ICASSP*, 20:645–648, 1998.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR 2015*, 2005.

[6] M. Dinarelli and S. Rosset. Models cascade for tree-structured named entity detection. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1269–1278, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

[7] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Proc. of Computer Vision and Pattern Recognition CVPR2010*, pages 2963–2970, 2010.

[8] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Face recognition from caption-based supervision. *IJCV*, 96(1), 2012.

[9] J.Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, 2005.

[10] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. pages 674–679, 1981.

[11] J. Luque, X. Anguera, A. Temko, and J. Hernando. Speaker diarization for conference room. *The UPC RT07s evaluation system. Multimodal Technologies for Perception of Humans.*, pages 543–553, 2008.

[12] J. Poignant, L. Besacier, G. Quenot, and F. Thollard. From text detection in videos to person identification. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 854–859, July 2012.

[13] J. Poignant, H. Bredin, and C. Barras. Multimodal person discovery in broadcast tv at mediaeval 2015. In *Proceedings of MediaEval 2015*, September 2015.

[14] P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, Nov. 1987.

[15] P. Salembier and L. Garrido. Binary partition tree as an efficient representation for image processing, segmentation and information retrieval. *IEEE TIP*, 9(4):561–575, April 2000.

[16] J. Shi and C. Tomasi. Good features to track. pages 593–600, 1994.

[17] R. Smith and G. Inc. An overview of the tesseract ocr engine. In *Proc. 9th IEEE Intl. Conf. on Document Analysis and Recognition (ICDAR*, pages 629–633, 2007.

[18] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, International Journal of Computer Vision, 1991.

[19] M. Uricar, V. Franc, and V. Hlavac. Facial landmarks detector learned by the structured output svm. In G. Csurka, M. Kraus, R. Laramee, P. Richard, and J. Braz, editors, *Computer Vision, Imaging and Computer Graphics. Theory and Application*, volume 359 of *Communications in Computer and Information Science*, pages 383–398. Springer Berlin Heidelberg, 2013.

[20] M. Zelenak and J. Hernando. The detection of overlapping speech with prosodic features for speaker diarization. *Proc. Interspeech*, 2011.

[21] M. Zelenak, C. Segura, J. Luque, and J. Hernando. Simultaneous speech detection with spatial features for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):436–446, 2012.