TokyoTech at MediaEval 2016 Multimodal Person Discovery in Broadcast TV Task

Fumito Nishi[1], Nakamasa Inoue[1], Koji Iwano[2], Koichi Shinoda[1]

[1]Tokyo Institute of Technology, Tokyo, Japan
[2]Tokyo City University, Kanagawa, Japan
{nishi, inoue, shinoda}@ks.cs.titech.ac.jp, iwano@tcu.ac.jp

ABSTRACT

This paper describes our diarization system for the Multimodal Person Discovery in Broadcast TV task of the MediaEval 2016 Benchmark evaluation campaign [1]. The goal of this task is naming speakers, who are appearing and speaking simultaneously in the video, without prior knowledge. Our diarization system relies on face diarization approach. We extract deep features from a face every 0.5 seconds, make visual i-vectors, cluster them, and associate results of clustering with optical character recognition.

1. INTRODUCTION

The Multimodal Person Discovery in Broadcast TV task can be split into subtasks: speaker diarization, face diarization, optical character recognition (OCR), speech transcription. We focus on diarization using face identification among these subtasks. This year, we introduce i-vectors [2] using deep features extracted from FaceNet, which is one of the state-of-the-art neural networks for face recognition.

Figure 1 shows the overview of our method. First, we detect and track faces in a video. Second, deep features are extracted from the detected faces at every 0.5 seconds. Third, i-vectors are made from deep features for each frame.

2. APPROACH

2.1 Face diarization

2.1.1 Deep feature

We employ FaceNet [3] to extract deep features. A deep feature is extracted from output layer of the network. It can measure similarity between faces. To extract deep features, face detection and tracking method in [4] is employed to obtain face regions in a video. Deep features are extracted from face regions at every 0.5 second.

2.1.2 Visual i-vectors

After deep features are extracted, we make visual i-vectors. I-vector is one of the state-of-the-art method for speaker verification. We apply this to the tracking segments.

Let M be a Gaussian Mixture Model (GMM) super-vector, which concatenates normalized mean vectors of an estimated GMM for a target video segment. An i-vector w is extracted from it, by assuming that M is modeled as

$$M = m + Tw,$$

where m is a face and channel independent super-vector, and T is a low rank matrix representing total variability. The Expectation Maximization (EM) algorithm is used to estimate the total variability as proposed in [2]. Note that w is associated with a given video segment. i-vector w_s for segment s is calculated by the following equation

$$w_s = (I + T^t \Sigma N(s) T)^{-1} T^t \Sigma^{-1} F(s),$$

where $N(s)$ and $F(s)$ are the zero, and first order Baum-Welch statistics on the Universal Background Model (UBM) for the current segment s, and Σ is the covariance matrix of the UBM. Each i-vector represents each face track respectively.

2.1.3 Attaching person's name tags

To attach a person's name tag for each face, the provided tags with time ranges obtained from optical character recognition (OCR) are used. First, each tag from OCR is attached to the face which has the maximum appearance time overlap. Here, we have tagged and untagged faces. Second, for each untagged face, we find the nearest tagged face to attach the same tag. If the distance between the untagged face and nearest tagged face is less than the predefined threshold, the same tag is attached to the untagged face. Distance between two faces are calculated by

$$D_{ij} = 1 - \frac{w_i w_j}{||w_i||_2 ||w_j||_2}$$

where w_i and w_j are i-vectors for tagged and untagged faces, respectively.

2.2 Speaker diarization

For speaker diarization, Bayesian Information Criterion (BIC) based segmentation with 12 MFCC + E is applied to obtain audio segments. Music and jingle segments are removed by Viterbi decoding. Finally, i-vectors are computed for each segment and clustered with Integer Linear Programming [5, 6].

2.3 Multimodal fusion

We employed the name propagation technique proposed in [7]. Our multimodal fusion takes intersection of tags obtained from speaker diarization and face diarization.

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

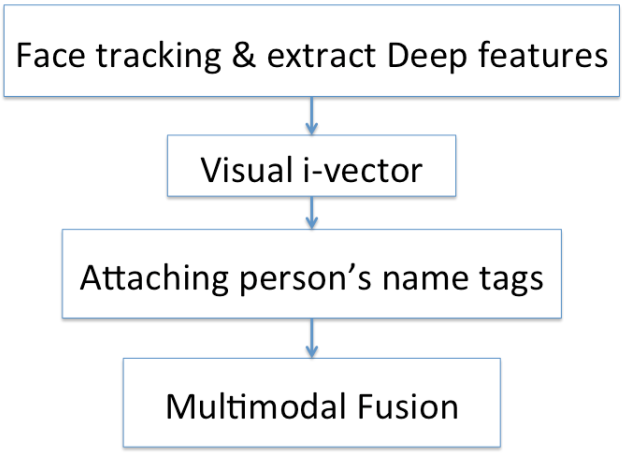

Figure 1: Overview of the whole system

Subset	Method	MAP@1 [%]	MAP@10 [%]
Leaderboard	Face Diarization (Primary)	27.7	18.3
	Face Diarization (Contrastive)	23.5	11.9
	Speaker Diarization	13.4	11.3
	Multimodal	22.7	15.0
Eval	Face Diarization (Primary)	31.5	20.0
	Face Diarization (Contrastive)	25.7	13.6
	Speaker Diarization	13.1	11.7
	Multimodal	29.3	17.3

Table 1: Mean Average Precision (MAP) in the test set

3. EXPERIMENTS AND RESULTS

3.1 Experimental Settings

We use dlib library [8] for face detection and tracking. For FaceNet, we use OpenFace implementation [9]. The dimension of deep features is 128. To extract visual i-vector, we trained UBM with 32 Gaussian mixtures and total-variability matrix on development set by using ALIZE [10]. The development set is the INA corpus which is used in the MediaEval 2015. We use detected faces for training. The dimension of visual i-vector is 100. For speaker diarization, we use the LIUM Speaker Diarization system [11]. The number of Gaussian mixtures for UBM is 256. The dimension of i-vector is 50. We used provided OCR and fusion code to build our system, and implemented all the other components.

3.2 Experimental Results

Table 1 shows MAP on the test set. Face Diarization is used for our submissions. The threshold used for the primary submission is adjusted in development set. The threshold of the contrastive submission is 0. Speaker Diarization and Multimodal evaluated by using the evaluation tool were not in our submission. Face Diarization is better than Speaker Diarization. It was effective for identifying speakers with short utterances. However, as we can see, Multimodal is worse than Face Diarization. To improve multimodal fusion system, we need to introduce multimodal features that can capture correlation between audio and visual streams. Modeling temporal relation between speakers is also needed to improve the performance.

4. CONCLUSION

We presented a face diarization based system, which uses visual i-vectors with FaceNet. Development of multimodal fusion methods and using sequential information is our future work.

5. REFERENCES

[1] Hervé Bredin, Camille Gauinaudeau, Claude Barras. Multimodal Person Discovery in Broadcast TV at MediaEval 2016. *Proc. of the MediaEval 2016 workshop*, Hilversum, Netherlands, Oct. 20-21, 2016.

[2] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798, 2011.

[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 886–893. IEEE, 2005.

[5] Mickael Rouvier and Sylvain Meignier. A global optimization framework for speaker diarization. In *Odyssey*, pp. 146–150, 2012.

[6] Grégor Dupuy, Sylvain Meignier, Paul Deléglise, and Yannick Esteve. Recent improvements on ilp-based clustering for broadcast news speaker diarization. In *Proceedings of Odyssey*. Citeseer, 2014.

[7] Johann Poignant, Hervé Bredin, Viet-Bac Le, Laurent Besacier, Claude Barras, and Georges Quénot. Unsupervised speaker identification using overlaid texts in tv broadcast. In *Interspeech 2012-Conference of the International Speech Communication Association*, p. 4p, 2012.

[8] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, Vol. 10, No. Jul, pp. 1755–1758, 2009.

[9] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[10] Bonastre, Jean-François and Wils, Frédéric and Meignier, Sylvain. ALIZE, a free toolkit for speaker recognition. *2005 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp.737–740, 2005.

[11] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier. An open-source state-of-the-art toolbox for broadcast news diarization. Technical report, Idiap, 2013.