LIG at MediaEval 2015 Multimodal Person Discovery in Broadcast TV Task

Mateusz Budnik, Bahjat Safadi, Laurent Besacier, Georges Quénot
Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
CNRS, LIG, F-38000 Grenoble, France
firstname.lastname@imag.fr

Ali Khodabakhsh, Cenk Demiroglu
Electrical and Computer Engineering Department
Ozyegin University, Istanbul, Turkey
ali.khodabakhsh@ozu.edu.tr
cenk.demiroglu@ozyegin.edu.tr

ABSTRACT

In this working notes paper the contribution of the LIG team (partnership between Univ. Grenoble Alpes and Ozyegin University) to the Multimodal Person Discovery in Broadcast TV task in MediaEval 2015 is presented. The task focused on unsupervised learning techniques. Two different approaches were submitted by the team. In the first one, new features for face and speech modalities were tested. In the second one, an alternative way to calculate the distance between face tracks and speech segments is presented. It also had a competitive MAP score and was able to beat the baseline.

1. INTRODUCTION

These working notes present the submissions proposed by LIG team (partnership between Univ. Grenoble Alpes and Ozyegin University) to the MediaEval 2015 Multimodal Person Discovery in Broadcast TV task. Along with the algorithms and initial results, a more general discussion about the task is provided as well. A detailed description of the task, the dataset, the evaluation metric and the baseline system can be found in the paper provided by the organizers [4]. All the approaches presented here are unsupervised (following the organizers guidelines) and were submitted to the main task.

The main goal of the task is to identify people appearing in various TV shows, mostly news or political debates. The task is limited to persons that speak and are visible at the same time (potential people of interest). Additionally, the task is confined to the multimodal data (including face, speech, overlaid text) found in the test set videos and is strictly unsupervised (no manual annotation available). The main source of names is given by the optical character recognition system used in the baseline [3].

Thanks to the provided baseline system [5], it was possible to concentrate on some aspects of the task, like a particular modality or the clustering method. Initially, our focus was on creating better face and speech descriptors. In the second approach however, only the distances between face tracks and speech segments were modified. The output of the baseline OCR system was used as is, while the output from the speech transcription system was not used at all.

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

2. APPROACH

Our initial approach focused on creating new features for both face and speech. The second approach is based more on the baseline system, i.e. no new descriptors were generated and the key element was the distance between speech segments and face tracks.

2.1 What did not work: new features

The first approach explored the use of alternative features for different modalities. For speech, a Total Variability Space (TVS) system [1] was designed using the following settings with the segmentation provided by the baseline system. Models were learned on the test data without any manual annotation available.

- 19 MFCC and energy + Δs (no static energy) + feature warping
- 20ms length window with a 10ms shift
- Energy based silence filtering
- 1024 GMMs + 400 dimensional TVS
- Cosine similarities between segments within each video are calculated

For faces, features extracted from a deep convolutional neural network [2] were used. This was done in the following way using the test set only:

- Face extraction with the approach provided by the organizers. All scaled to resolution of 100×100 pixels.
- Labels generated by the OCR. They were then assigned to co-occurring faces. This was based on a temporal overlap between the face and the label. This generated list served as a training set. The number of classes equaled the number of unique names.
- The general structure of the net is based on the smallest architecture presented in [6], but with just 5 convolutional layers and the number of filters at each layer reduced by half. The fully connected layers had 1024 outputs. It was trained for around 15 epochs.
- After the training, the last layer containing the classes was discarded and the last fully connected hidden layer (1024 outputs) was then used for feature extraction.

Two individual sets of clusters were generated for each modality. Afterwards, both were mapped to the shots. If there was an overlap with the same label, the person was named. Additional submissions involving this approach were made, which included adding descriptors provided by the baseline (e.g. HOG for face and BIC for speech). However, they did not manage to give better performance than the baseline.

2.2 What did work : modified distance between modalities

In the baseline provided, the written names are first propagated to speaker cluster and then the named speakers are assigned to co-occurring faces. Due to the nature of the test set, an alternative was used where the written names are first propagated to face clusters. These face-name pairs are subsequently assigned to co-occurring speech segments. This approach yielded a more precise but smaller set of named people compared to the baseline. In order to expand it, a fusion with the output of the baseline system was made, where every conflict (e.g. different names for the same shot) would be resolved in favor of our proposed approach.

Additionally, another way to calculate the distance between speech and face track was developed. In the baseline the distance between a face track and a speech segment is calculated using lip movement detection, size and the position of the face and so on. Our complementary approach is based on temporal correlation of tracks from different modalities.

First, overlapping face tracks and speech segments are extracted for each video. Similarity vectors for both modalities are extracted with respect to all the other segments within the same video. Correlation of the similarity vectors are calculated in order to determine which face and voice go together. In other words, a face-speech pair which appears frequently throughout the video is more likely to belong to the same person. Finally, the output of this approach is fused with the output of the system described in the first paragraph of this subsection (face-name pairs assigned to co-occuring speech segments) to produce a single name for each shot.

3. INITIAL RESULTS AND DISCUSSION

The first system (submitted for the first deadline) performed rather poorly with 30.48 % EwMAP (MAP = 30.63 %). While our second approach, submitted as the main system by the second deadline, with EwMAP = 85.67 % (and MAP = 86.03 %) was far more successful and was able to beat the baseline system (EwMAP = 78.35 % and MAP = 78.64 %). The scores presented here were provided by the organizers and can change slightly before the workshop, due to more annotation being available.

During the preparation for this evaluation there were a number of issues and observations connected to both our approach and to the data. First of all, trying to build biometric models for individual people does not work well for this particular task (at least based on what was tested in the context of this evaluation, e.g. SVMs). In order to comply with the task requirements, the labels can only be generated from the OCR and then be assigned to one of the modalities. However, both steps are unsupervised, generating noisy annotation in the process. Additionally, the video test set consists of one type of program (TV news) where, apart from the news anchor, most people appear only once and this may not be enough to create an accurate biometric model. This stands in contrast to the development set, which contains debates and parliament sessions where some persons re-appeared much more frequently.

A more general issue is also the class imbalance. While some people, especially the anchors, appear frequently across different videos, most of the others are shown once or twice and are confined to a single video. This makes the use of unsupervised techniques, like clustering, challenging, due to widely varying cluster sizes - small clusters can get attached to bigger ones, which is heavily penalized under the MAP metric. This can, at least partially, explain the poor performance of the first approach. Even though the features used in this method are state-of-the-art, they would require more high quality data (including annotation) and parameter adjustment to create good enough distinctions between thousands of individual persons appearing in the videos.

4. CONCLUSIONS

During this evaluation different algorithms were tested in order to (unsupervisingly) identify people, which speak and are visible in TV broadcasts. One approach concentrated on trying to provide state-of-the-art features for different modalities, while the other provided an alternative estimation of the distance between already provided modalities of face and speech.

The first approach, even with its limited performance on this particular shared task, seems to have greater potential and our future work may try to address some of its shortcomings. This includes a focus on a more robust deep learning approach that could deal with noisy or automatically generated training sets.

5. ACKNOWLEDGMENTS

This work was conducted as a part of the CHIST-ERA CAMOMILE project, which was funded by the ANR (Agence Nationale de la Recherche, France).

6. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

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

[4] J. Poignant, H. Bredin, and C. Barras. Multimodal person discovery in broadcast tv at mediaeval 2015. *MediaEval 2015 Workshop*, September 2015.

[5] J. Poignant, H. Bredin, V. Le, L. Besacier, C. Barras, and G. Quénot. Unsupervised speaker identification using overlaid texts in tv broadcast. *INTERSPEECH*, 2012.

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.