IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

The GTM-UVIGO System for Audiovisual Diarization

Eduardo Ramos-Muguerza, Laura Docío-Fernández, José Luis Alba-Castro

AtlanTTic Research Center, University of Vigo

eramos, ldocio, jalba@gts.uvigo.es

Abstract

This paper explains in detail the Audiovisual system deployed by the Multimedia Technologies Group (GTM) of the atlanTTic research center at the University of Vigo, for the Albayzin Multimodal Diarization Challenge (MDC) organized in the Iberspeech 2018 conference. This system is characterized by the use of state of the art face and speaker verification embeddings trained with publicly available Deep Neural Networks. Video and audio tracks are processed separately to obtain a matrix of confidence values of each time segment that are finally fused to make joint decisions on the speaker diarization result.

Index Terms: speaker recognition, face recognition, deep neural networks, image processing.

1. Introduction

In recent years, the field of pattern recognition has witnessed a shift from the extraction of handmade features to machine-learned features using complex neural network models. Biometric verification is a clear example of an application scenario where Deep Neural Networks have produced a notable increase in performance, providing transformations of space where the face and voice of users are represented in clusters that are more compact and separable than in the original sample space. This representation makes the problem of diarization and verification in multimedia content more tractable than with previous approaches [1][2][3][4].

However, facial and speaker verification models are still not perfect and make many mistakes in verifying the identity of people in natural conditions. These situations are common when analyzing audiovisual content with constant shot changes and different types of scenarios, variability in the appearance of faces (pose, expression and size), variability in the mix of voices, noise and background music. Also, the appearance of many other people who are not registered to be identified and are considered "intruders" to the system, causes many false identity assignments.

In this paper we explain the approach that the GTM research group has followed to tackle the person identification problem in audiovisual content. We have prepared a system that works separately on the video and audio tracks and makes a final fusion to fine tune the speaker diarization result. The rest of the paper is organized as follows. Section 2 explains the video processing part, including the segmentation of the video footage into different shots and the face detection, tracking, verification and post-processing at shot level. Section 3 explains the Speaker Diarization and Verification subsystem. Section 4 deals with the fusion of modalities and Section 5 gives the computational cost information. Finally, section 6 presents the conclusions and details the on-going research lines.

2. Video Processing

Television programs such as news, debates, interviews, etc., are characterized by frequent changes of shot and scene, the appearance of multiple people and the mixture of different scenes in the final configuration of frame that consumes the end user. In this way, the final audiovisual content is very different from the typical scenarios where biometric identification is used, such as restricted access, video security or mobile scenarios.

The solution we have adopted for this competition in the video processing part is based on two fundamental ideas that apply to this type of content. On the one hand, we know that a change of shot implies, in general, a change in the person who appears on the scene, although it does not always happen and it does not happen in the same way regarding the speaker. On the other hand, the people who appear in a shot remain in it as long as there is no movement of the camera or of the people themselves. This way, detection of shot changes gives an important clue for subsequent face processing.

2.1. Detection of shot changes

This subsection explains a simple approach to detect shot changes that is designed to have more false positives than false negatives. Shot changes will be used to restart face trackers because we cannot rely on tracking a face through shot changes, so losing a shot change could have a greater impact in the tracker than initializing the face tracker unnecessarily.

Detection of movement is also an important feature to have a more complete understanding of the footage, but we haven't included in this version of the system a specific movement detection block. Instead, we have used the false positive rate of the shot detection block as an indication of movement.

The steps to detect a change of shot are the following:

1. Reduce the size of the frame to save computational load,
2. Calculate the derivatives of the image to keep the edges of the scene,
3. Divide the frame into blocks and calculate the mean of edge pixels per block,
4. Subtract the mean of the same block in the previous frame,
5. Set a threshold for considering that a block difference represents a change (threshold set with the development video footage),
6. Count the number of block changes and set a threshold defined for a change of shot (also using the development set).

This approach is very simple and fast. It also leaves a lot of room for improvement by defining areas with different shot change thresholds depending on the type of scene or type of video realization. For example, in some of the videos of the competition, the consumed scene have the frame divided into

10.21437/IberSPEECH.2018-41

different areas with different video content. It is quite common that the shots change at different pace in the different areas, so a solution that can make local decisions on shot change is quite useful. However, we left for future improvements of the system the local detection of shot changes. For this version we just set a permissive global threshold that allows detection of total or partial change of shot, movement and fading as a unique event.

2.2. Face processing

The face processing subsystem comprises several sequential operations that are briefly explained through the Figure 1 and in the subsections below.

2.2.1. Face Detection and Geometric Normalization

Face detection is a fundamental step in the sequential processing. We have used the detector based on Multi-Task Cascaded Convolutional neural Network [5], that jointly finds a Bounding Box for the face and five landmarking points useful to normalize the face. This face detector is quite robust to pose, expression and illumination changes. False negatives are typical in extreme poses with yaw angles beyond +/-60º and pitch angles beyond +/- 40º, that are not so uncommon in interview and debate contents. This approach also brings a bit amount of false positives in areas where textured objects with skin colors appear, like hands, arms and other not human objects.

Once a face is detected (being true detection or not), its bounding box (BB) is saved with several parameters that will allow to do tracking and assign identities during the process. An overlapping function between the current BB, and the BBs of the previous frame allows linking the BBs belonging to the same person and do backtracking when the shot has finished.

The detected face is passed to a geometric normalization that prepares the face to be plugged in in a standardized way to the face recognition block.

2.2.2. Face recognition

We have used the face recognizer based on dlib's implementation [6] of the Microsoft ResNet DNN [1]. This DNN finds an embedded space where similar faces are grouped together and far from different faces. This network is also trained to be quite robust to pose, expression and illumination changes. In this case, the network makes quite many false identification assignments when poses are beyond +/-50º in yaw and +/- 20º in pitch. Also extreme facial expressions produce false assignments, being quite robust, though, for neutral and smiling faces (the great bulk of face images found in internet-based datasets for face recognition). However, TV contents offer more facial expressions of emotion than neutral and smiling, so false assignments are quite common also in these cases.

After a face is detected for the first time in a shot, meaning that no previous BB is linked to the current one, a candidate ID is assigned to the BB if it surpasses a distance threshold for new IDs (Th_newID) when comparing the embedded vector against all the embedded vectors of the enrollment set. The closest ID is kept for that BB. A confidence value is also assigned to that ID in that specific BB. If the BB is linked to a previous BB, the embedded vector is compared just with the cluster of enrolled vectors of the previous candidate ID. If it surpasses a second threshold (Th_previousID) then the same ID is assigned to the BB. The rationale of keeping two different thresholds and having Th_newID > Th_previousID is that assigning an ID to a new detection should be more restrictive than assigning it to a previous one located in an overlapped area. It is important to highlight that a shot change can make a face with different ID to appear in the same position that another face in the previous frame before the shot. So, a change of shot restarts the tracking process.

2.2.3. Face clusters augmentation

The unreliable robustness of the face recognizer with extreme poses and expressions lead us to define a strategy to cope with potential erroneous ID assignments.

A previous study of the capacity of dlib's facial recognition network, advised us to use a restrictive Th_newID, so comparison of faces with extreme poses between the current embedding and the embeddings in the enrollment dataset will not surpass it. This way, it is less likely that a wrong ID assignment is propagated through a tracked BB. Only when the head moves to a more frontal pose, or a similar extreme pose is stored in the enrollment dataset, the ID is assigned to the BB and propagated. But, what happens if the face in the shot is always in an extreme pose? We need a way to enrich the enrollment dataset with new samples of a specific ID that appear in the content but are different enough from the enrollment dataset. Given that the Th_previousID is more relaxed, during a shot where a previous match has been assigned, several different face samples with almost any pose and expression can enrich the enrollment dataset. The criterion to enrich the dataset is the increase of variance of the ID face cluster. Then, the enriched dataset is ready to use in the next frame.

2.2.4. Face ID backtracking

Once a shot change is detected, an online post-processing is run to reassign IDs or even delete potentially wrong assignments in the past shot. This block is based on heuristic rules defined after observing the typical behavior of the previous processing blocks in the development scenarios. If the detector and recognizer were ideal, a shot without rapid movements or with slow camera movement should contain just a number of BB tracks that matches the number of persons in the shot; the first BB should appear in the first shot frame and the last BB of its BB track should appear in the last frame of the shot. But things are not perfect, so the rationale of the backtracking post-processing is based on the next observations:

- False negatives of the detector break the paths of the BBs, so there will be more paths than in the ideal case.
- False initial matchings in every new track will yield tracks with incorrect IDs.
- The number of persons in the shot can be roughly estimated by the average number of detections. This assumption is broken when extreme poses appear during a long period, producing an underestimated number of persons in the shot.
- Every ID matched in the shot can appear in several broken paths. The accumulated confidence of that ID gives a rough approximation of its probability of being in that shot.

Using these observations we defined a rule to keep tracks of a number of persons just above the estimated average and with the IDs with largest accumulated confidence. Those IDs are finally assigned to the time intervals within and across shots and written down in an rttm file.

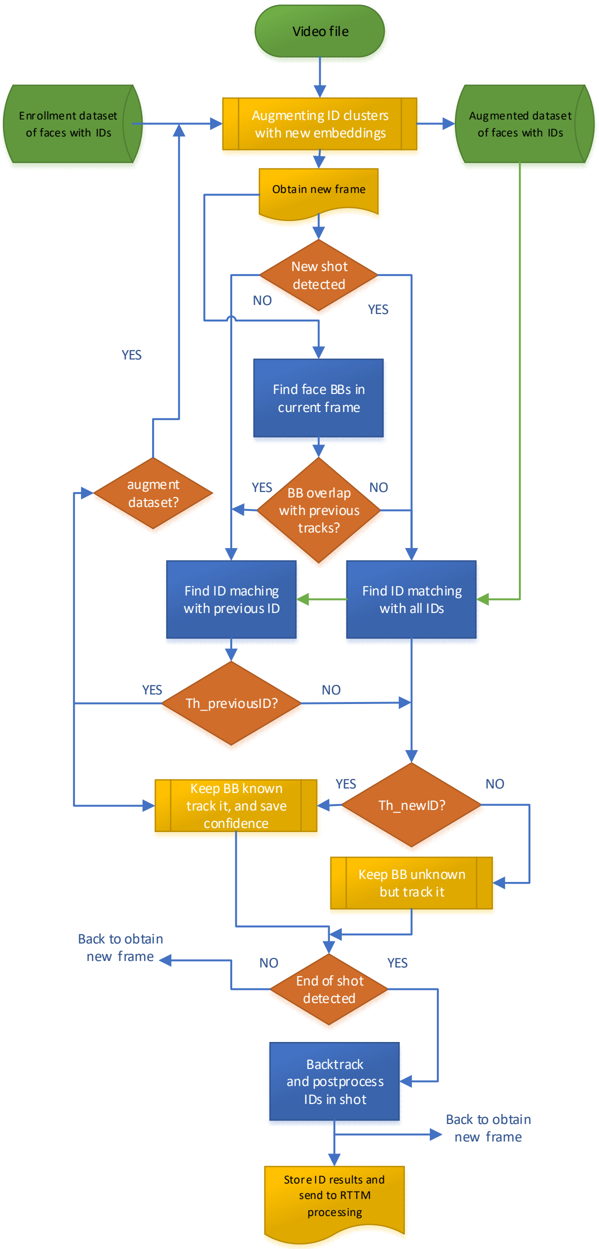
Video file
Enrollment dataset of faces with IDs
Augmenting ID clusters with new embeddings
Augmented dataset of faces with IDs
Obtain new frame
New shot detected
NO
YES
YES
Find face BBs in current frame
augment dataset?
YES
BB overlap with previous tracks?
NO
Find ID maching with previous ID
Find ID matching with all IDs
YES
Th_previousID?
NO
Keep BB known track it, and save confidence
YES
Th_newID?
NO
Keep BB unknown but track it
Back to obtain new frame
NO
End of shot detected
YES
Backtrack and postprocess IDs in shot
Back to obtain new frame
Store ID results and send to RTTM processing

Figure 1: *Flow diagram of face processing.*

3. Speaker Diarization and Recognition

The developed strategy for speaker diarization and verification uses a DNN trained to discriminate between speakers, and which maps variable-length utterances or speech segments to fixed-dimensional embeddings that are also called x-vectors [2].

A pretrained deep neural network downloaded from http://kaldi-asr.org/models.html was used. The network was implemented using the nnet3 neural network library in the Kaldi Speech Recognition Toolkit [7] and trained on augmented VoxCeleb 1 and VoxCeleb 2 data [8].

Figure 2 represents the block diagram of the speaker diarization and recognition subsystem.

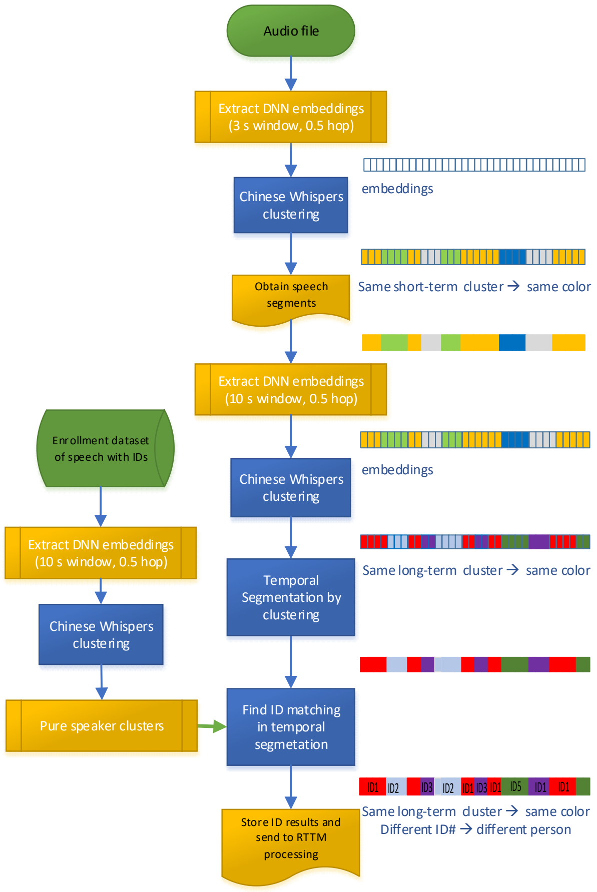
Audio file
Extract DNN embeddings (3 s window, 0.5 hop)
embeddings
Chinese Whispers clustering
Obtain speech segments
Same short-term cluster → same color
Extract DNN embeddings (10 s window, 0.5 hop)
Enrollment dataset of speech with IDs
embeddings
Chinese Whispers clustering
Extract DNN embeddings (10 s window, 0.5 hop)
Same long-term cluster → same color
Temporal Segmentation by clustering
Chinese Whispers clustering
Pure speaker clusters
Find ID matching in temporal segmetation
ID1 ID2 ID3 ID2 ID1 ID3 ID1 ID5 ID1 ID1
Same long-term cluster → same color
Different ID# → different person
Store ID results and send to RTTM processing

Figure 2: *Block diagram of the speaker diarization and recognition subsystem.*

3.1. Speaker enrollment

The audio signal provided for each person in the enrollment set is used to obtain DNN speech-based embeddings. A sliding window of at least 10 seconds with a half a second hop is used. Then, these embeddings are clustered using the Chinese Whispers algorithm [9]. The threshold of the clustering algorithm is adjusted so that the clusters are pure and at least as many as the number of identities in the enrollment set. In this way an enrolled person can be represented by one or more clusters.

3.2. Off-line Speaker Diarization

First, the audio signal is divided into 3 second segments with a half a second hop. DNN short-term audio embeddings were extracted for each of these segment, clustered using the Chinese Whispers algorithm and their timestamps kept. From the clustering result we obtain an audio segmentation. Next, each of these segments, of arbitrary duration, are processed in order to extract one or more long-term audio embeddings using the same DNN. To do this, a sliding window of at least 10 seconds with a half a second hop is used. Then, these embeddings are clustered using again the Chinese Whispers algorithm, using a threshold that minimizes the diarization error.

3.3. On-line Identity Assignment

The clusters obtained in the previous step need to be assigned to the enrollment identities. Keeping the timestamps of each embedding in the clustering process, allows to design an online ID assignment approach. Temporal segments are defined as consecutive timestamps with embeddings associated to the same cluster. The ID assigned to a time segment is the enrollment ID of the best-matching enrollment cluster, as far as this distance is less than a threshold. This threshold is defined after observing the typical behavior of the system in the development scenarios. A confidence value for that ID in that specific temporal segment is stored to be used jointly with the face-based confidence value in the fusion process.

4. Fusion

Once a decision was made for both modalities, a multimodal fusion approach was implemented in order to correct potentially wrong speech-based ID assignments. Given a temporal segment that has been assigned a speaker identity ID1, if a high-confidence single face identity ID2 has been detected in more than 60% of the video frames in that speaker ID1 segment, the speaker ID1 is changed to the face identity ID2.

This rule doesn't apply if ID1 and ID2 have different gender (as given by the enrollment name).

Much more elaborated fusion rules can be applied at this stage, but only this one was tested for the competition.

The results over the Development video provided by the organizers of the competition are presented in Table 1.

Table 1: *DER results on the Development video*

Modality	DER
Face	36.20%
Speaker	14.25%
Speaker Fusion	7.51%
Average DER	21.85%

5. Computational Cost

The computational cost of the proposed audiovisual diarization system was measured in terms of the real-time factor (RT). This measure represents the amount of time needed to process one second of audiovisual content: xRT = P/I, where I is the duration of the processed video and P is the time required for processing it. The whole development video was processed to compute the RT, thus taking into account many different audiovisual situations. The duration of this video is I = 7410 s, and the time needed to process it was P = 33457 s, leading to RT = 4.51. These computation time was obtained by running this experiment on an Intel(R) Core(™) i5 CPU 670@3.47 GHz with 12 GB RAM. Even though the process is running more than 4 times slower than real-time, the code is not optimized at all (some parts are coded in Matlab) and the machine is just using 1 CPU and no GPU. We are working to speed up the process and expect to have it running in real-time in the next months.

6. Conclusions and future work

We have presented the GTM-UVIGO System deployed for the Albayzin Multimodal Diarization Competition at Iberspeech 2018. The system uses state of the art DNN algorithms for face detection and verification and also for speaker diarization and verification. The application scenario is studied to implement ad-hoc post-processing strategies to fine-tune the ID assignments made by the video and audio parts. Specifically, the information on shot changes are exploited to avoid tracking faces across shots. Confidence matrices are used in a fusion strategy that allows changing pre-assigned speaker identities.

This framework leaves a lot of room for improvement in each of the fundamental processing stages and also in the ad-hoc rules for fine-tuning. One of the main future lines consist of increasing the robustness of face matchings for extreme poses and expressions and the robustness of speaker ID assignments in overlapped speech. From a video processing point of view, a better characterization of the display montage would allow the application of post-processing rules less prone to errors. Finally, speeding up some DNN critical parts by using GPUs and efficiently coding some other parts would allow real-time processing of the system.

Acknowledgements

This work has received financial support from the Spanish Ministerio de Economía y Competitividad through project 'TraceThem' TEC2015-65345-P, from the Xunta de Galicia (Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019) and the European Union (European Regional Development Fund ERDF).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH 2017 – 97th Annual Conference of the International Speech Communication Association*, Proceedings, pp. 999–1003, 2017.

[3] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5115–5119.

[4] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey and A. McCree, "Speaker diarization using deep neural network embeddings," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 4930-4934.

[5] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, Oct. 2016.

[6] D. E. King. "Dlib-ml: A Machine Learning Toolkit, Journal of Machine Learning Research," 10:1755-1758, 2009.

[7] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 1-42011.

[8] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large scale speaker identification dataset," *INTERSPEECH 2017 – 97th Annual Conference of the International Speech Communication Association*, 2017.

[9] Chris Biemann, "Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems," in *First Workshop on Graph Based Methods for Natural Language Processing (TextGraphs-1)*, pp. 73-80, 2006.