IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

The GTM-UVIGO System for Audiovisual Diarization 2020

Manuel Porta-Lorenzo, José Luis Alba-Castro, Laura Docío-Fernández

AtlanTTic Research Center, University of Vigo
mporta, jalba, ldocio@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 2020 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 and fine-tuned for the persons of interest. Video and audio tracks are processed separately and are finally fused to make joint decisions on the speaker diarization result. Few modifications have been made over the GTM-UVIGO system presented in the very same conference in 2018, mainly regarding the video processing part.

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

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 frequent shot changes, camera movement, different types of scenarios, variability in the appearance of faces (pose, expression, illumination, blurring and small 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. Sections 4 and 5 present the experimental results Finally, Section 6 gives the computational cost information, and Section 7 presents the conclusions and details the on-going research lines.

2. Face diarization

Television programs such as news, debates, interviews, documentaries, etc., are characterized by frequent changes of shot and scene, the appearance of multiple people in foreground and also people and dynamic content in the background. On the other hand, other programs like TV-series can contain faces in very extreme appearances regarding pose, expression, illumination, make-up and size. Therefore, 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 in the scene, although it does not always happen and it does not happen in the same way regarding the speaker or the type of program. 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. In the updated system for the 2020 challenge, television programs are much more diverse and the assumption that the camera is still most of the time in each shot does not hold anymore. Also, some of the programs have large digital screens at the background showing dynamic content that can mislead the former module of shot change detection. So, in the next section we explain the main changes applied to this module to cope with the high false detection rate produced by camera or background dynamics.

2.1. Detection of shot changes

This subsection explains a simple approach to detect shot changes designed to work in a ROC point with FP > FN. 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,

10.21437/IberSPEECH.2021-17

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 an upper and a lower threshold, defined for reliable changes of shot and shot continuity, respectively (also using the development set).
7. For the cases with values between both thresholds, detect all the faces in the frame and compare them with the detections of the previous frame in terms of position, size and distance between faces against a third threshold.

This "Canny-style" thresholding allows the system to track faces over smooth transitions and camera movements, potentially procuring longer tracks which are more suitable for the shot-based face processing. Nevertheless, these thresholds are quite dependent on the type of program and video realization, and so it is left for future improvements of the system the dynamic adaptation of those thresholds via the detection of different kinds of program. For this version a set of permissive values which minimize the number of false negatives for the whole ensemble of target videos was empirically obtained.

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, and quite typical in TV-series. 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 processing full tracks 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. Training and fine-tuning a face recognizer

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 faces with the same ID are grouped together and are far from faces with different ID.

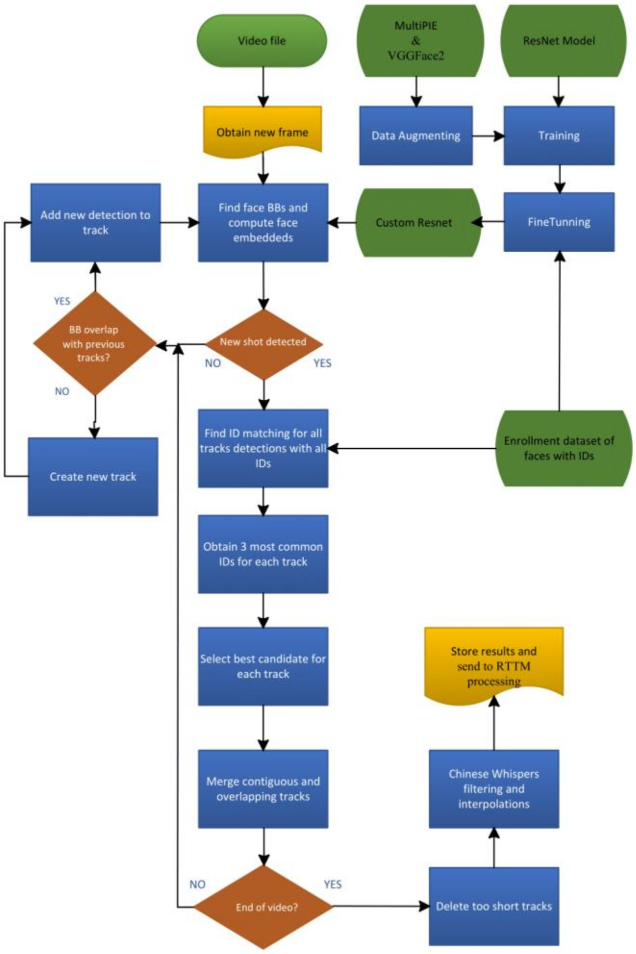
Video file
MultiPIE & VGGFace2
ResNet Model
Obtain new frame
Data Augmenting
Training
Add new detection to track
Find face BBs and compute face embeddeds
Custom Resnet
FineTunning
YES
BB overlap with previous tracks?
NO
New shot detected
NO
YES
Create new track
Find ID matching for all tracks detections with all IDs
Enrollment dataset of faces with IDs
Obtain 3 most common IDs for each track
Select best candidate for each track
Store results and send to RTTM processing
Merge contiguous and overlapping tracks
Chinese Whispers filtering and interpolations
NO
YES
End of video?
Delete too short tracks

Figure 1. *Flow diagram of face processing.*

In order to improve the performance of the original dlib's network when poses are beyond +/-50º in yaw and +/- 20º in pitch, and face expressions are not neutral or smiling (the most typical in public datasets), we have retrained the network from scratch using the VGGFace2 [7] dataset and a subset of the MultiPIE face database [8] that contains extreme poses, not neutral expressions and non-uniform illumination. Re-scaling was also performed for augmenting small face samples. This network was then fine-tuned to pay more attention to the faces of the enrollment set. For the 2020 competition, 161 identities were provided, with 10 images per ID. Hand reviewing and normalization of incorrect aspect-ratio of many images was necessary to avoid the network learning aspect-ratios that would not appear in the training, development and test videos. Data augmentation was applied by horizontal flipping, XY-shifting, and downsampling-upsampling to provide blurred images. With this augmented image set a fine tuning of the previous network was programmed with the next characteristics:

- All weights of the DNN are allowed to learn through SGD with an initial learning rate of 1e-6. This decision worked better than freezing weights for some convolutional layers and allowing learning to the rest and the FC final layer.

- In order to avoid the catastrophic forgetting effect when fine tuning the Face recognizer, we have resorted to the rehearsal technique [9] and added face samples from the original training with VGGFace2 dataset. Not doing so, would make the model tending to memorize the enrollment IDs but failing more on unknown IDs, increasing False Alarm rate. Mini Batches of 100 identities and 10 samples per identity were randomly but evenly extracted from the augmented enrollment dataset and the VGGFace2 dataset, and trained with a cross-validation stop condition over the LFW (Labeled Faces in the Wild) dataset [14]. Fine tuning stops as soon as the FAR+FRR decreases a cumulative 1% over 5 steps (5000 face images). This method ensures that the face recognizer is tuned to the enrollment set but doesn't forget to tell apart also unknown IDs. It is important to note that the VGGFace2 dataset was reduced by taking out the LFW identities.

2.2.3. Shot-based face processing

When a face is detected in a shot its Bounding Box is compared with the last element of the current shot tracks and it is assigned to the closer one if it surpass a distance threshold or to a new one otherwise (or if there aren't any tracks on the current shot). As a shot change can make two faces belonging to different IDs to appear in the same position, a change of shot restarts the tracking process.

After a change of shot is detected, meaning that each stored track accumulates samples of the same person, a candidate ID is assigned to each track by comparing the embedded vectors of all its samples with all the embedded vectors of the enrollment set. For each sample, the closest enrollment ID is obtained and the three most frequent ones among all samples are taken into account. From these three candidates, the one which achieves the smallest individual distance between any pair of track and enrollment set samples is selected as the candidate ID and such distance is stored as the track reference value. Afterwards, in order to filter false positives and faces from persons who don't belong to the enrollment set, the reference value of the track is compared against a dynamic threshold which is defined to be less restrictive for the IDs with higher presence in the video so far. The rationale behind this method is to make the most of the tracking information, using the higher quality detections to select the candidate ID for all the samples in a track, and to ease the recognition of the persons having more presence in the video. Once a track has an assigned ID it is compared to the previously processed tracks of the same shot in terms of assigned ID, distance between bounding boxes and separation in time in order to detect fragments belonging to the same ID and join them by merging its detections. The processed tracks are stored and the program jumps to the next shot until the end of the video.

When the full video is processed the obtained tracks pass through different filters to enhance the performance of the system. First, the tracks which contain less than a fixed number of frames are deleted as they mostly correspond to detection errors. Following, the embeddings of each track are clustered using a Chinese Whispers algorithm [12] keeping only those belonging to the main cluster. Finally, the frame indexes and bounding boxes of each track's detections are interpolated to obtain continuous segments.

3. Speaker diarization

The used strategy for speaker diarization and verification is similar to those of the GTM-UVIGO system presented in the 2018 Challenge [13]. Specifically, it 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 [10] and trained on augmented VoxCeleb 1 [11] and VoxCeleb 2 data [15].

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 [12]. 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. Time 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 behaviour of the system in the development scenarios. A confidence value for that ID in that specific time segment is stored to be used jointly with the face-based confidence value in the fusion process.

3.4. Fusion

The SPEAKER modality in the 2020 contest with 161 IDs of enrollment, produced too many false assignments of time segments. To correct potentially wrong speech-based ID assignments, a multimodal fusion approach that uses the assignments made by both modalities separately was implemented. Given a time segment that has been assigned a speaker identity ID1, three rules are applied depending on the FACE modality content:

1. If no faces were assigned to any enrollment ID in the same time segment, and also, the identity ID1 is not found anywhere in the video using the face modality, the speaker ID1 is removed from the speaker output file. That is, it is very unlikely that when someone speaks, no face will appear in the video in that time interval and, furthermore, that identity will not appear in the whole video.
2. If a high-confidence single face identity ID2 has been detected in more than 60% of the video frames in that speaker ID1 time segment, and the identity ID1 is not found anywhere in the video using the face modality, the speaker ID1 is changed to the face identity ID2. That is, it is very unlikely that someone will speak on the video and never appear his face.
3. If several face identities (including ID1) were assigned with high-confidence in the same time segment that the speaker identity ID1, and a single face identity ID2 has been detected in enough frames (above the 60%), the assigned 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). This makes the face modality more reliable than the speaker one.

4. Results on Development videos

The primary evaluation metric to rank systems is the average of the face and speaker diarization errors (Averaged DER). Performance metrics over one of the Development videos provided by the organizers of the competition are presented in Table 1. It is worth noting that we consider this video as a testing one, that is, the enrollment identities to search were those of the Test dataset.

Table 1: *Results on one of the Development videos*

Modality	DER	MISSED	FALARM	ERROR
Face (F)	15.77	11.9	3.8	0
Speaker (S)	32.01	9.2	20.3	2.5
Speaker Fusion (SF)	24.81	9.6	15.2	0
Averaged F & SF	20.29	10.75	9.5	0

5. Results on Tests videos

The results over the Test videos provided by the organizers of the competition are presented in Figure 2. In this graph, the speaker DER refers to the modified output after fusion with Face output. X axis shows the acronym for the program and Y-axis the DER (face - BLUE, speaker fusion - RED and average of both - ORANGE). The system meta-parameters have been adapted to the details of the AT ("Aquí la Tierra") program used for developing, so it seems clear the dependency of the system to the type of program, especially in the SPEAKER modality, but also in FACE. It is also noticeable the bad performance of SPEAKER compared to FACE in the SFT ("Si Fueras Tu") program.

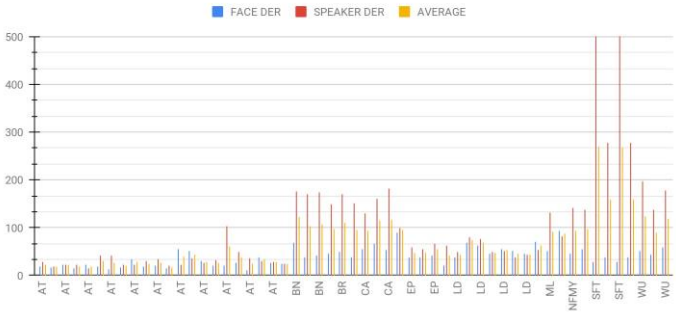

Figure 2. *DER results on Test videos*

6. 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. An example video (AT-20181111.mp4) was processed to compute the RT, thus taking into account many different audiovisual situations. The duration of this video is I = 1743.72 s, and the time needed to process it was P = 7229.5 s, leading to RT = 4.146. This computation time was obtained by running this experiment on an Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz with 256 GB RAM. Even though the process is running more than 4 times slower than real-time, the code is not optimized at all (it is completely implemented in Python and the machine is not fully exploited.

7. Conclusions and futures work

We have presented the GTM-UVIGO System deployed for the Albayzin Multimodal Diarization Competition at Iberspeech 2020. The system uses state of the art DNN algorithms for face detection and verification and also for speaker diarization and verification including fine tuning of the face recognition model using the persons of interest. In order to minimize tracking errors, a shot detection algorithm resets the face trackers and makes use of a Canny-style double threshold to optimize the change decision. A novel shot-based faces tracking is also proposed, which makes the most of the temporal information, retrieves low quality faces, filters false positives using a dynamic threshold and provides continuity to the output detections. 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. This framework leaves a lot of room for improvement in each of the fundamental processing stages and also in the ad-hoc rules for post shot fine-tuning.

8. Acknowledgements

This work has received financial support from the Xunta de Galicia (Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019; AtlanTTic and ED431B 2018/60 grants) and the European Union (European Regional Development Fund ERDF).

9. 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] Cao, Qiong, Shen, Li, Xie, Weidi, Parkhi, Omkar and Zisserman, Andrew. (2018). VGGFace2: A Dataset for Recognising Faces across Pose and Age. 67-74. 10.1109/FG.2018.00020.

[8] http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

[9] Robins, Anthony. "Catastrophic Forgetting, Rehearsal and Pseudorehearsal". Connection Science. 7 (2): 123–146.1995

[10] 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.

[11] 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.

[12] 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.

[13] Eduardo Ramos-Muguerza, L. Docío-Fernández and J. L. Alba-Castro. “The GTM-UVIGO System for Audiovisual Diarization”, In Proceedings of the IberSPEECH, 2018; pp. 204–207.

[14] G.B. Huang, M. Ramesh, T. Berg and E. Learned-Miller. “Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments”, University of Massachusetts, Amherst, Technical Report 07-49, 2007.

[15] J. S. Chung, A. Nagrani, A. Zisserman. “VoxCeleb2: Deep Speaker Recognition”, INTERSPEECH, 2018.