Who is really talking? A Visual-based Speaker Diarization Strategy

Pedro A. Marín-Reyes[1], Javier Lorenzo-Navarro[1] Modesto Castrillón-Santana[1], and Elena Sánchez-Nielsen[2]

[1] Instituto Universitario SIANI, 35017, Las Palmas, Spain,
Universidad de las Palmas de Gran Canaria
[2] Departamento de Ingeniería Informática y de Sistemas, 38271, Santa Cruz de Tenerife, Spain,
Universidad de la Laguna
pedro.marin102@alu.ulpgc.es

Abstract. The speaker activity at the Canary Islands Parliament is recorded, and later manually annotated. This task can be modelled as a diarization problem, that is a way to automatically annotated who and when is speaking. In this paper, we propose the use of the visual cue to solve the diarization task. To perform this approach, it is mandatory to detect individuals, determine the one speaking, and extract features for matching. In order to test the performance of our proposal, we evaluate four different strategies based on the visual shot features.

Keywords: visual diarization strategies, local descriptors, histogram distances, F-reid

1 Introduction

Speaker Diarization deals with annotating who and when a speaker is talking, it represents a challenge for the scientific community [1, 2] that is mostly tackled using the audio cue. This problem can be tackled from a vision-based point of view, considering a re-identification process, i.e. detecting a speaker and checking whether he/she appears again.

A standard solution to audio-based speaker diarization is based on the procedure described by Tranter and Reynolds [3], being the approach adopted by the most recent literature. The purpose of speaker diarization is to split the audio recording of the different people interventions into segments. In this way, each segment represents a single speaker. After that, a clustering technique is used to group the different segments in order to include all the segments of one person in the same cluster. Different diarization scenarios have captured the attention of researchers, specially of those who investigate in the field of audio signals.

Ning et al. [4] have focused on Japanese Parliament sessions to the aim to solve speaker diarization, they segment the speech using Mel Frequency Cepstral Coefficient (MFCC) and Bayesian Information Criterion (BIC) as features. Then, the Kullback-Leibler (KL) divergence is used at the clustering process as

Fig. 1. Different views of the Canary Islands Parliament.

similarity measure between segments, obtaining the number of clusters by the value of the eigenvalues of the affinity matrix. These techniques are also used by Lupu et al. [5] in the Rumanian Parliament, using the system LIUM [6] to extract the audio of the sessions without taking into account the visual information of the videos.

To improve the results of only audio methodologies, Campr et al. [7] proposed the use of audio and visual information applied to Czech parliamentary recordings. Using Gaussian Mixture Model (GMM) to segment and detect in the audio any new speaker or recorded, for the latter also update the parameters of the corresponding GMM. After the face is detected and normalized, Local Binary Pattern features are extracted. For each group of consecutive faces, a cluster of key-faces are selected, to be later matched with different clusters, and they are compared among the different clusters. If the distance between two clusters is lower than a threshold, they are considered to be the same identity, otherwise it is a new person. After that, using the fusion of both diarization processes, the number of models is reduced with the audio-based diarization.

Furthermore, video processing can be used to detect the speakers, even without the audio information. Everingham et al. [8] propose a method to automatic annotation of film characters. To this purpose, both the subtitles and facial information are analized, where Scale-Invariant Feature Transform (SIFT) descriptor is used, and the clothing characterized by the YCbCr color histogram. In some cases, a person who is not speaking appears in the image, so, a speaker detector is implemented using the consecutive histogram differences of the mouth area. The matching process is done by a distance scheme of each character with the nearest representation of the face and clothing to assign an identity. Then, a Support Vector Machine (SVM) classifier is trained, one class with respect the others. Unlike the previous work, Sang and Xu [9] use scripts, instead of the

subtitles to identify the name of the speaker. When all the faces are detected, they are grouped into several clusters using a clustering technique, matching the face identify using a graph fit, Error Correcting Graph Matching (ECGM).

The contributions of this paper are the following: 1) propose different strategies to assign the speaker ID from visual segments to an audio segment, 2) study different local descriptors to apply the above assignation, and 3) compare different distances to measure the similarity between descriptors in the problem of assign an ID.

2 Scenario

In this paper, we are focused on the diarization of parliamentary debates sessions using only video information. Specifically, this work is based on the Canary Islands Parliament in Spain. In this scenario, speaker interventions can be done from three different points: 1) at the presidential table where presidential deputies follow the guidelines to expose the topic during a predefined speaking time, 2) at the platform located, at front of the presidential table where the deputies explain some topics, and 3) at the seats, the place where the deputies are sitting, in some cases they can stand up and intervene to answer another deputy. In those places, the interventions are recorded by a network of cameras distributed in the Parliament, which can do pan, tilt and zoom. Fig. 1 shows different images recorded in the Parliament. Those cameras are managed by a producer who decide the camera to focus the attention, which could lead to change the view during the intervention of a speaker, that situation increases the problem challenges involved in a vision-based system because the camera could be recording a person who is not talking.

3 Procedure

The speaker diarization problem is tackled based on a visual approach, using the face as the main source of information. For each detected face in each frame, the following processing is applied: Initially, the image is rotated till the position of the eyes is horizontal. Then, to generate the model, the faces have to satisfy the Biometric Keyframe condition [10], where the eyes and mouth distances match with a frontal pose.

Later, each key-face is transformed to grey-scale and the face or head shoulder (HS) pattern are obtained as a region of interest (ROI); the face pattern represents a ROI of the face area, and the HS pattern is composed by the face area adding the surrounding information as hair, clothing and background are included. Then, as features, local descriptors are extracted to obtain an histogram representation, using different grid size setups, because they have demonstrated good performance in facial analysis [11], an outline of the process is shown in Fig. 2. After the ROI is modelled, a matching stage is carried out by comparing the model against the database models, that is updated with each new identity

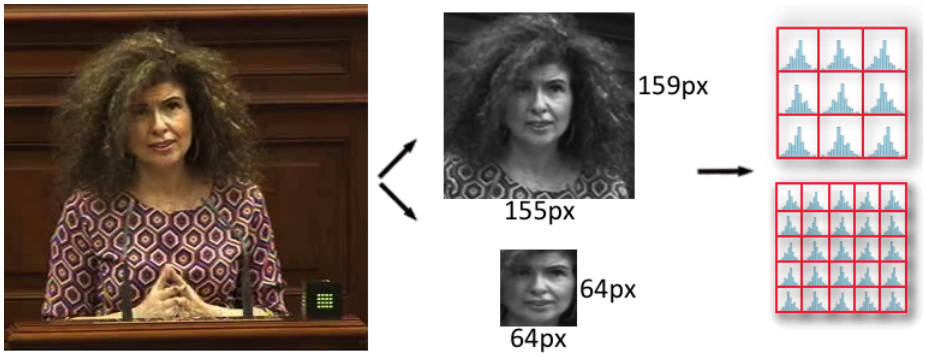

Fig. 2. The image is normalized using face or HS pattern. Then, it is divided into 3×3 or 5×5 grids respectively where a local descriptor is applied to obtain the speaker model.

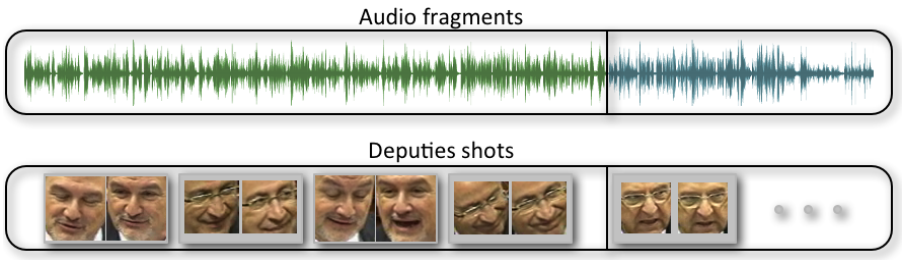

Fig. 3. Audio fragments can include different visual shots.

found as the video is processed. The comparison is made using a histogram distance. The minimum distance model of the database is taken, and if this distance is larger than a fixed threshold, it is considered as a new speaker and added to the database, otherwise it is the same identity.

Besides, the parliamentary sessions, as well as other debates scenarios, we have to deal with shot changes while the intervener is talking as commented in section 2. Therefore during a speaker intervention, i.e. to annotate his/her audio fragment, different deputies shots could appear, see an example in Fig. 3. The system has to give to the whole audio fragment corresponding to a single speaker the same ID through different visual shots. If the assignment is not correct, the diarization system will annotate with a wrong ID this shot or in worst case, the system creates a new ID, and for future comparisons the system will take into account the new false speaker. That increases the number of failures to the re-identification task. To solve the above problem, different strategies are proposed:

– **First Appearance (FA)**: The person of the first shot detected by the system in the audio fragment is taken as representative.

– **Most Frequent (MF)**: The person that the system detects a larger number of times in the audio fragment is taken as representative speaker shot.
– **Greatest Length (GL)**: The person that the system detects as largest duration shot in the audio fragment is taken as representative speaker shot.
– **Greatest Total Length (GTL)**: The person that the system detects as largest duration in the audio fragment is taken as representative speaker shot.

4 Experiments

The experiments have been performed using 29 videos[3] of different sessions of the Parliament. Those videos present different number of frames, shots and speakers; the mean duration of the videos is four hours. As local descriptors we have considered to test the following ones:

– Histogram of Oriented Gradients (HOG)
– Local Binary Patterns (LBP)
– LBP Uniform (LBPu2)
– Intensity based LBP (NILBP)
– Local Gradient Patterns (LGP)
– Local Phase Quantization (LPQ)
– Local Salient Patterns (LSP0)
– Local Ternary Patterns (LTP)
– Local Oriented Statistics Information Booster (LOSIB)
– LTP high (LTPh)
– LTP low (LTPl)
– Weber Local Descriptor (WLD)

Local descriptors are calculated in the mentioned ROI areas(face or HS) using a 5×5 and 3×3 grids. At the same time, the comparison of the models is computed with Canberra, Chebyshev, Cosine, Euclidean and kullback-Leibler divergence histogram measures. Besides, the different configurations are computed using the diarization strategies commented in the previous section.

To get an idea of the cost of the carried out experiments, 29 videos with two patterns with 2 grid configurations with 12 local descriptors with 5 different measures were processed, making a total of 6,960 experiments. Moreover, those experiments were validated by four diarization approximations, obtaining a total of 27,840 experiments.

4.1 Results

To evaluate the results, True Re-identification Rate (TRR) and True Distinction Rate (TDR) are taken from [12]. The TRR measure determinates how good is the system to re-identificate individuals, and the TDR represents the measure

[3] Videos available at http://www.parcan.es

of how good is the system to distinguish between individuals. At the time, to evaluate the system, it could be the case that a system assigns only different IDs to the individuals detected, it will obtain 0% in TRR and 100% in TDR. We need to combine those values, at first, we will take the mean value. But, it will obtain a 50% being the worst system possible. To avoid this problem, it is taken the F_1 score, labelled as F_{reid}, that combines the TRR and TDR measures, as it is shown in Eq. 1.

$$F_{reid} = 2 \cdot \frac{TRR \cdot TDR}{TRR + TDR} \tag{1}$$

To focus in diarization methods we have calculate the F_{reid} mean value of all the videos processed. Although, the mean value of the different local descriptors and distance measures, see Table 1, where the Most Frequent approach matches the highest value independently of the kind of ROI and grid configuration. Taking into account this setup, the results improve 2.61% in the best case.

Strategy	Face		HS	
	3x3	5x5	3x3	5x5
FA	56.61	52.23	64.03	59.51
MF	56.70	54.91	**64.31**	59.57
GL	56.19	54.22	63.85	59.05
GTL	54.21	54.65	61.70	57.65

Table 1. Comparison of different patterns and number of grid respect different diarization approaches in term of the F_{reid} for the mean value of all the videos processed, descriptors and distances.

Table 2 shows the comparison of different local descriptors with different pattern and grid configuration. The best descriptor is Weber Local Descriptor that obtains an increment of 0.39% in relation to the second best descriptor, Histogram Oriented Gradients. The former obtains an improvement of 5.86% in relation to the worst descriptor for this configuration.

In relation to the histogram distance, Canberra is the best distance that matches the highest value, as we can see in Table 3. But in general, Kullback-Leibler divergence has a good behaviour for the different configurations and the difference between Canberra and this measure is insignificant.

Additionally, we highlight the use of HS and a 3×3 cells division, that obtains the best results for all the experiments. Specifically, the best configuration reported 74.09% with the Most Frequent approach, using a HS pattern with a 3×3 grid applying Weber Local Descriptor and comparing the models with a Canberra distance.

Descriptor	Face		HS	
	3x3	5x5	3x3	5x5
HOG	56.81	55.09	65.86	63.51
LBP	55.03	53.17	62.10	58.36
LBPu2	55.95	55.52	63.93	58.97
LGP	53.52	51.72	64.58	59.41
LOSIB	49.56	47.57	64.72	60.65
LPQ	58.75	53.19	60.62	55.65
LSP0	55.49	54.70	59.25	53.75
LTPh	56.87	54.35	62.42	58.79
LTPl	56.40	54.05	63.29	57.02
LTP	56.97	54.65	62.75	59.77
NILBP	56.60	56.66	65.91	57.63
WLD	59.20	57.34	**66.25**	63.83

Table 2. Comparison of different patterns and number of grids respect different local descriptors in term of the F_{reid} for the mean value of all the videos processed, diarization approaches and distances.

Distance	Face		HS	
	3x3	5x5	3x3	5x5
Canberra	54.13	53.53	**65.86**	62.16
Chebyshev	52.84	47.25	55.76	46.81
Cosine	57.58	55.94	65.04	62.50
Euclidean	58.74	55.86	65.16	60.07
KL	56.35	57.43	65.54	63.18

Table 3. Comparison of different patterns and number of grids respect different histogram distance measures in term of the F_{reid} for the mean value of all the videos processed, diarization approaches and descriptors.

5 Conclusion

This paper addresses four different strategies related to diarization problems, where these strategies avoid the annotation of false speaker in a vision-based context. Furthermore, the purpose of this article is to test various features related to computer vision to obtain a good configuration of parameters. So, Different local descriptors have been compared using HS and face patterns with two grid configurations, obtaining a general idea of their behaviour. Finally, multiple histogram measures have been compared, allowing us to know what configuration give us greater results for upcoming test.

In general, HS pattern matches the best results independently of the other parameters. In a same way, the 3×3 grid increases the performance of our diarization system. Moreover, Weber Local Descriptor is the best form to reduce the dimensionality of our problem, getting good results. At the time to compare the models, Canberra reports the best values. And last but not less important, to use a Most Frequent approach in a diarization system avoids the apparition of false speakers identification with an increment of 2.61% in terms of F_{reid}

comparing the different diarization approaches using the HS pattern with a 3×3 grid.

Acknowledgement

This work is partially supported by Government of Spain through TIN2015-64395-R and by the Ministerio de Economía y Competitividad, Government of Spain and FEDER funds of the European Union through TIN2016-78919-R (MINECO/FEDER).

References

1. Miró, X.A., Bozonnet, S., Evans, N.W.D., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech & Language Processing **20**(2) (2012) 356–370
2. Barra-Chicote, R., Pardo, J.M., Ferreiros, J., Montero, J.M.: Speaker diarization based on intensity channel contribution. IEEE Transactions on Audio, Speech & Language Processing **19**(4) (2011) 754–761
3. Tranter, S.E., Reynolds, D.A.: An overview of automatic speaker diarization systems. IEEE Transactions on Audio, Speech, and Language Processing **14**(5) (Sept 2006) 1557–1565
4. Ning, H., Liu, M., Tang, H., Huang, T.: A spectral clustering approach to speaker diarization. In: Proc. ICSLP. (2006)
5. Lupu, E., Apatean, A., Arsinte, R.: Speaker diarization experiments for romanian parliamentary speech. In: 2015 International Symposium on Signals, Circuits and Systems (ISSCS). (July 2015) 1–4
6. Meignier, S., Merlin, T.: Lium spkdiarization: an open source toolkit for diarization. In: CMU SPUD Workshop, Dallas (Texas, USA) (mars 2010)
7. Campr, P., Kunešová, M., Vaněk, J., Čech, J., Psutka, J. In: Audio-Video Speaker Diarization for Unsupervised Speaker and Face Model Creation. Springer International Publishing, Cham (2014) 465–472
8. Everingham, M., Sivic, J., Zisserman, A.: Taking the bite out of automated naming of characters in tv video. Image and Vision Computing **27**(5) (April 2009) 545–559
9. Sang, J., Xu, C.: Robust face-name graph matching for movie character identification. IEEE Transactions on Multimedia **14**(3) (June 2012) 586–596
10. Marín-Reyes, P.A., Lorenzo-Navarro, J., Castrillón-Santana, M., Sánchez-Nielsen, E.: Shot classification and keyframe detection for vision based speakers diarization in parliamentary debates. In: Conference of the Spanish Association for Artificial Intelligence, Springer (2016) 48–57
11. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E.: Multi-scale score level fusion of local descriptors for gender classification in the wild. Multimedia Tools and Applications **(in press)** (2016)
12. Cong, D.N.T., Khoudour, L., Achard, C., Meurie, C., Lezoray, O.: People re-identification by spectral classification of silhouettes. Signal Processing **90**(8) (2010) 2362 – 2374 Special Section on Processing and Analysis of High-Dimensional Masses of Image and Signal Data.