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

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. A way to automatically annotate what speaker is talking now, it could be done using a visual diarization process, that it takes place in the Canary Islands Parliament where the pppintervenes are recorded to realize a list with the different apparitions of the deputies lately. To perform this task, it is mandatory to detect and normalize the deputies to obtain a model. Although, those model have to be compared to identify what speaker is. In addition, it is necessary to identify if a intervener detection is talking, to match good results we propose four approaches based on the visual shot features.

Keywords: Local descriptors, visual diarization, F-reid

1 Introduction

Speaker Diarization deals with annotating who and when a speaker is talking, it represents a challenge for scientific community [1]. This problem can be tackled from a point of view of a re-identification process, detecting a speaker and checking whether he/she appears again.

A possible solution to speaker diarization could be done using the procedure of the author [2], being the approach adopted by the most recent literature. The purpose of speaker diarization is to split the audio recording of the different interventions of the people into segments, in this way, each segment represent an unique speaker. After that, it is used a clustering technique to group the different segment in order to have all the segment of one person in each cluster. Diarization problems have captured the attention of researches, specially of those who investigate in the field of audio signals.

Ning et al. [3] have focus on Japanese Parliament sessions to the aim of solve speaker diarization, the speech is segmented using MFCC (Mel Frequency Cepstral Coefficient) and BIC (Bayesian Information Criterion) as features. Then, the KL (Kullback-Leibler) divergence is used at the clustering process as similarity measure between segments, obtaining the cluster number by the value of

the eigenvalues of the affinity matrix. Theses techniques also are used by [4] in the Rumanian Parliament using the system LIUM [5] to extract the audio of the sessions without take account the visual information of the parliament videos.

To improve the results of only audio methodologies, Campr et all. [6] proposed the use of audio and the visual information in Czech parliamentary recordings. Using GMM (Gaussian Mixture Model) to segment and detect in the audio the new speaker or recorded for later update the parameters of the corresponding GMM. Face detection and tracking are used to the corresponding video processing. After the face is normalized, it is extracted the LBP (Local Binary Patterns) features. Of each group of consecutive faces it is selected a cluster of key-faces and they are compared between the different clusters. If the distance between clusters is lower than a threshold, it is consider to be the same identity or it is a new person. After that, using the fusion of both diarization process, it is reduced the number of models that are obtained with the audio-based diarization.

Video process can be used to detect the speakers, even without the audio information. Everingham et al. [7] propose a method to automatic annotate the film characters. To this purpose, it is used the information of the subtitles as well the face visual information, where Scale-Invariant Feature Transform (SIFT) descriptor is obtained and the clothing Characterized by the YCbCr colour histogram. In some cases, the character that do not speak 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, SVM classifier is trained, one class respect the others. Unlike the previous work, Sang and Xu [8] use scripts, instead of the subtitles to identify the name of the speaker. When all the faces are detected, they are group using a clustering technique into several clusters like speakers and then it is proposed the identity matching of the faces using graph fit, Error Correcting Graph Matching (ECGM).

2 Scenario

In this paper, we are interested in diarization of parliamentary debates sessions based only on video. Specifically, this work is focus on the the Canary Islands Parliament. The deputy interventions can be done at the presidential table where the presidential deputies follows the guidelines to expose the topic and the speaking time of each deputy; at the platform, it is the place at front of the presidential table where the deputies explain some topics; and at the seats, it is the place where the deputies are sitting, in some cases, a deputy 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, in Fig. 1 shows an example of the images recorded in the Parliament. Those cameras are managed by a producer group who decide the camera to focus the attention, which could lead to change the view during the intervention of a person, that situation generates a biggest problem to a

Fig. 1. Different views of the Parliament.

vision-based system because the camera could be recording a person who is not talking.

3 Procedure and Experiments

In this paper, the speaker diarization is based using a visual approach, it is focus on the face information. For Each face detection of the speaker, it has to be processed, at first time, the image is rotated till the position of the eyes where horizontal. To generate the model, the faces have to satisfy the Biometric Keyframe condition [9] where it is processed the position of the eyes and the mouth to get the faces that have approximately a frontal pose.

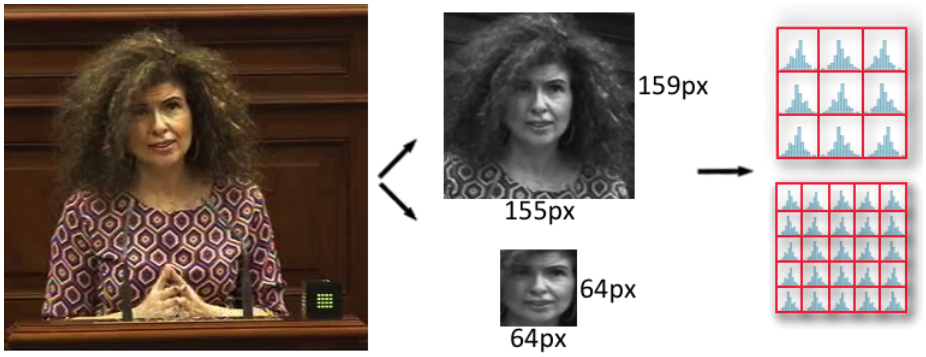

Fig. 2. The image is normalized using face or HS pattern. Then, it is divided in 3×3 or 5×5 grids where a local descriptor is applied to obtain the speaker model.

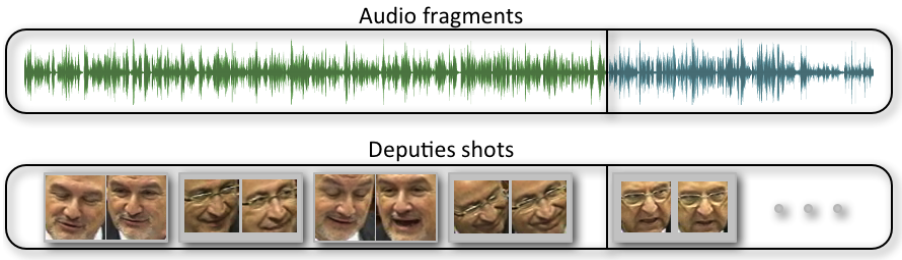

Fig. 3. Audio fragments can include different visual shots.

Each key face selected have to transform the image space colour to grey scale and the face or HS pattern are obtained as region of interest (ROI) and then, as features, local descriptors are extracted using different size of grid because they have demonstrated good performance in facial analysis [10], an example of the process is shown in Fig. 2. After the ROI is modelled, a matching stage is carried out by comparing the extracted features against the database models. This comparison is made using a histogram distance, the minimum distance model of the database is taken and if this distance is upper than a determine threshold, it is considered as a new speaker or instead, it is the same.

The experiments have been realized using 29 videos[3]. Those videos variate in number of frames, shots and speakers. The mean duration of the videos is three hours and a half. As local descriptors we have considered in the comparison the following ones: Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), LBP Uniform (LBPu2), Intensity based LBP (NI-LBP), Local Gradient Patterns (LGP), Local Phase Quantization (LPQ), Local Salient Patterns (LSP0), Local Ternary Patterns (LTP), LTP high (LTPh), LTP low (LTPl), Weber Local Descriptor (WLD) and Local Oriented Statistics Information Booster (LOSIB). They are calculated in the ROI area using a 5×5 and 3×3 grids. At the time, the comparison of the models were calculated with Canberra, Chebyshev, Cosine, Euclidean and kullbackLeibler (KL) divergence histogram measures.

Besides, in visual diarization problems we have to face the change of shot to another deputy while the intervener is talking as it is commented in section 2. So, for a specific speaker audio annotation, audio fragment, could appear different deputies shots in the video sequences, see an example in Fig. 3. For this reason, we propose the following four diarization approaches:

- **First Appearance (FA)**: The person of the first shot detected by the system in the audio fragment is taken as representative speaker shot.
- **Most Frequent (MF)**: The person that the system detect as greater number of occurrences in the audio fragment is taken as representative speaker shot.
- **Greatest Length (GL)**: The person that the system detect as higher duration shot in the audio fragment is taken as representative speaker shot.

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

– **Greatest Total Length (GTL)**: The person that the system detect as higher duration in the audio fragment is taken as representative speaker shot.

To get an idea of the cost of experiments executed, 29 videos with two patterns with two kinds of grids with 12 local descriptors with five different measures were processed, a total of 6,960 experiments. Moreover, those experiments were validated by four diarization approximations, obtaining a total of 27,840 experiments.

4 Results

To evaluate the results, True Re-identification Rate (TRR) and True Distinction Rate (TDR) are taken by [11]. The TRR measure determinates how good are the system to re-identificate individuals and the TDR represent the measure 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 assign only different ID 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 as F_{reid}, that combines the TRR and TDR measures, as it is shown in Eq. 1.

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

4.1 Diarization approaches

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 Most Frequent match the highest value independently of the kind of ROI and the grid configuration. Taking into account this approach, we increase the results in 2.61% in the best case.

Approach	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 pattern 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 respect the pattern and grids. The best descriptor is the Weber Local Descriptor that obtains an increment of 0.39% respect the second best descriptor, Histogram Oriented Gradients, and an improvement of 5.86% respect the worst descriptor for this configuration.

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

At the time to choose one histogram distance we have to take care because those distances influence to much in the results, as we can see in Table 3. Canberra is the best distance that match the highest value. But in general, kullback-Leibler divergence have a good behaviour for the different configurations and the difference between Canberra and this measure is insignificant.

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 pattern and number of grid respect different histogram distance measures in term of the F_{reid} for the mean value of all the videos processed, diarization approaches and descriptors.

Although, it has obtain the best result using the HS and a nine cells division in general. Summarizing the results, it have been obtained 74.09% the best

configuration with the Most Frequent approach, using Weber Local Descriptor in a 3×3 grid in a HS pattern and comparing the models with a Canberra distance.

5 Conclusion

To summarize this paper, four approaches to measure the performance of diarization problems have been proposed. Moreover, different local descriptors were compared using HS and face patterns with two grids configuration, obtaining a general idea of their behaviour. Finally, multiple histograms measures to matching have been compared, allowing us to know what configuration give us greater results for upcoming test.

The purpose of this article is to test various features related to computer vision to obtain a good configuration of parameters. HS pattern in general match the best results independently of the other parameters. In a same way, 3×3 grid increases the performance of our diarization system. Moreover, WLD is the best form to reduce the dimensionality of our problem, getting good results. At the time to compare the models, Canberra scores 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 a3×3 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.

References

1. 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
2. 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
3. Ning, H., Liu, M., Tang, H., Huang, T.: A spectral clustering approach to speaker diarization. In: Proc. ICSLP. (2006)
4. 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
5. Meignier, S., Merlin, T.: Lium spkdiarization: an open source toolkit for diarization. In: CMU SPUD Workshop, Dallas (Texas, USA) (mars 2010)

6. 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
7. 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
8. Sang, J., Xu, C.: Robust face-name graph matching for movie character identification. IEEE Transactions on Multimedia **14**(3) (June 2012) 586–596
9. 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
10. 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)
11. 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.