Video Indexing Using Face Appearance and Shot Transition Detection

Dario Cazzato
Interdisciplinary Center for Security, Reliability and Trust (SnT)
University of Luxembourg
dario.cazzato@uni.lu

Marco Leo
Institute of Applied Sciences and Intelligent Systems
National Research Council of Italy
marco.leo@cnr.it

Pierluigi Carcagnì
Institute of Applied Sciences and Intelligent Systems
National Research Council of Italy
pierluigi.carcagni@cnr.it

Cosimo Distante
Institute of Applied Sciences and Intelligent Systems
National Research Council of Italy
cosimo.distante@cnr.it

Javier Lorenzo-Navarro
Instituto Universitario SIANI
Universidad de las Palmas de Gran Canaria
javier.lorenzo@ulpgc.es

Holger Voos
Interdisciplinary Center for Security, Reliability and Trust (SnT)
University of Luxembourg
holger.voos@uni.lu

Abstract

The possibility to automatically index human faces in videos could lead to a wide range of applications such as automatic video content analysis, data mining, on-demand streaming, etc. Most relevant works in the literature gather full indexing of videos in real scenarios by exploiting additional media features (e.g. audio and text) that are fused with facial appearance information to make the whole frameworks accurate and robust. Anyway, there exist some application contexts where multimedia data are either not available or reliable and for which available solutions are not well suited. This paper tries to explore this challenging research path by introducing a new fully computer vision based video indexing pipeline. The system has been validated and tested in two different typical scenarios where no-multimedia data could be exploited: broadcasted political video documentaries and healthcare therapies sessions about non-verbal skills.

1. Introduction

Videos represent one of the most important media and the automatic analysis of their contents is one of the most investigated topics in computer science related research areas.

Since human face is an important subject in videos, because it is a unique feature of human beings and is ubiquitous, most of proposed video analysis strategies rely on a match of facial features. Thanks to deep learning based approaches a very high accuracy in face matching and retrieval has been reached even on very large image database [25]. Although images in datasets were acquired with different facial expressions, illumination conditions and occlusions, the exploitation of the above mentioned approaches in real world scenarios is not trivial. The most impressive application works deal with movie actor indexing, but it is worth noting that the prior knowledge of appearance models for a cast [13] and the number of searching clusters (i.e. the number of characters) makes the task easier [1]. Besides, support by existing video captions [27] or other contextual knowledge is required [3]. In addition to the particular application field just mentioned, it is quite common in the literature to find works gathering full indexing of videos in real scenarios by exploiting additional media features (audio [15, 24] and text [12, 8]) that are fused with facial appearance information to make the whole frameworks accurate and robust. Also the task of speaker diarization in videos has taken benefit from state-of-the-art face clustering [4], spatiotemporal Bayesian fusion [14] and Fisher Linear Discriminant analysis [31] for both audio and video signals.

Anyway, there exist some application contexts where

multimedia data are either not available or reliable. Two relevant examples are: 1) the indexing of people in news videos, documentaries or reportage (where audio is not necessarily correlated with images due to the presence of the narrator) and 2) video assisted healthcare (where different subjects, often with cognitive diseases, appear sequentially in front of the camera and their non-verbal behaviours are recorded). These cases require a fully appearance-based video indexing but, unfortunately, to the best of our knowledge, this is a very few investigated research topic.

This paper tries to explore this challenging research path by introducing a new computer vision based video indexing pipeline. The proposed pipeline can detect and re-identify people in each image supplying an automatic subject dependent video indexing that can be exploited for different purposes, e.g. to monitor balanced participation in politic talk shows and news, or to handle electronic records in healthcare systems more easily. It works without using any prior knowledge about the number of relevant persons in the video. Moreover, faces are processed without any constraints about appearance, eyeglasses, beard or hairstyle. A shot transition detection system is also introduced and opportunely integrated in the pipeline in order to improve the results. The system has been validated and tested in two different scenarios. At first, the method has been employed for indexing politicians framed in videos of the Canary Islands Parliament (Spain) sessions. In addition, it was exploited for indexing children in videos acquired during sessions aimed at assessing Autism Spectrum Disorders.

The rest if the paper is organized as follows: relevant works on face indexing are reported in Section 2, while the proposed method is presented in details in Section 3. Experiments and results are showed and discussed in 4; Section 5 concludes the paper with conclusions and future works.

2. Related Work

The problem of re-identifying the same person under multiple views and different perspectives was initially addressed as a multi-camera tracking problem [19]. For a complete review and the evolution of re-identification works, refer to [38]. The person re-identification became a new and independent research line starting from the work in [16], in which the overall appearance of the individual was specifically modelled to deal with several challenges such as different camera angles and illumination conditions, variation in pose and the rapidly changing appearance of loose or wrinkled clothing. Since that pioneering approach, a plethora of works with the same underlying idea has been proposed. In [11], a segmentation model is used to separate the foreground to whom a computer vision pipeline is applied in order to obtain multi-shot re-identification. In [32], Dynamic Time Warping (DTW), widely used for action recognition, has been employed to re-identify the person. A top-push distance learning model (TDL) has been proposed to enforce the optimization on top-rank matching in re-identification combined with the minimization of intra-class variations to improve performance [35]. Wang *et al.* [34] introduce a discriminative video ranking model by simultaneously selecting reliable space-time features from video fragments. 3D Histogram of Oriented Gradients (HOG3D) features and optical flow energy profile over image sequence are used to produce a representation designed to generate multiple fragments from unregulated video clips. AdaBoost is applied to both Haar-like features and dominant colour descriptors to achieve the most invariant and discriminative signature in [9].

Several approaches have also been proposed in order to achieve efficient video indexing and retrieval like keyframes texture, edge and motion features [29], temporal mapping [2], or temporal patterns combined with sequence matching techniques [21]. Faces have been automatically annotated by clustering methods with a weighted feature fusion in [7], dealing with colour information, but a training set is needed in order to perform a general-learning (GL) training scheme. Blind separation, i.e. labelling with lack of any prior knowledge, was proposed in [28], but the method worked well only in case of a limited number of participants, relative stable video scene and face images captured in frontal-view. Different clustering metrics have also been exploited to fully automatic indexing people when the total number of people is unknown [6].

Some related works relying only on image analysis limit their scope to the identification of main characters' [17] without achieving the indexing of videos, or scalable face retrieval over large datasets [26]. In [20], the face appearance is joint with body information to enhance the face recognition on videos. Deep learning approaches for face retrieval have also recently been proposed [10, 23, 33, 37].

3. Proposed Method

The proposed method works as follows: first of all, the video is processed in order to detect the different shots by a frame analysis based on the Kullback-Leibler (KL) divergence. In particular, for each shot, information about first, last frames and duration in frames and seconds is stored. During the same video scanning, facial images are extracted and tracked by a framework of cascaded convolutional neural networks (CNNs). Facial Features are then extracted by a ResNet-50 pre-trained on the VGGFace2 dataset [5] and the cosine distance among feature vectors is computed and used as a metric to guess a match among face belonging to the same person. In order to reduce errors, if faces belongs to the same shot, then threshold is relaxed, whereas a strongest threshold is used in case of matching in different camera shots. A block diagram of the proposed solution is reported in Figure 1. In the next subsections, each algorith-

mic step will be detailed.

3.1. Shot Transition Detection

Video segmentation into shots allows to characterize them in terms of boundaries and duration, in frames and seconds. The segmentation of the video is carried out following the method proposed by Sánchez-Nielsen *et al.* [30]. The method is based on the computation of the Kullback-Leibler divergence (D_{KL}) between every two consecutive frames, $frame_i$ and $frame_{i+1}$, to assess how similar are their color distribution, P_i and P_{i+1} respectively. The color distribution of each frame is computed as the concatenation of the YCbCr components histogram discretized into 16 bins and concatenated to get one histogram per frame. After that, a normalization stage is done. Under the hypothesis that two frames of the same shot have similar color distributions, shot boundaries are detected when the KL divergence between the color distribution of two consecutive frames $D_{KL}(P_i||P_{i+1})$ is larger than a threshold. Thus,

$$frame_i = \begin{cases} \text{boundary} & if D_{KL}(P_i||P_{i+1}) > thr \\ \text{not boundary} & otherwise \end{cases} \quad (1)$$

In this work the threshold value, thr, has been set to 0.05.

3.2. Multi-Face detection

Since faces can vary in appearance for changes in pose, lighting and occlusions, face detection is a challenging task in real-world applications. In this paper, the results recently proposed in [36] have been used. The aforementioned work relies on a framework of three multi-task cascaded CNNs. In our work, the input image is initially resized to different scales to build an image pyramid which is the input of the cascaded framework. The same fully convolutional network of [36] called Proposal Network (P-Net) is exploited to obtain the candidate windows. Non-maximum suppression (NMS) is then used to collapse highly overlapping boxes. Face localization is further improved by using a bounding box regression module. All candidates are fed to another CNN, called Refine Network (R-Net), which rejects false candidates. Differently from [36], the filtered information of face classification is only extracted and then further processed in our algorithmic pipeline. The configurations of the two CNNs are reported in Figure 2.

3.3. Facial Feature Extraction

The employed network architecture in this step is the SE-ResNet-50-256D with 30M parameters trained on the VGGFace2 dataset [5]. The considered architecture contains a the "Squeeze-and-Excitation" (SE) block, that adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels [18]. The VGGFace2 dataset contains 3.31 million images of 9131 subjects (identities), with an average of 362.6 images for each subject. Images are downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity and profession (e.g. actors, athletes, politicians). The whole dataset is split to a training set (including 8631 identites) and a test set (including 500 identites). A face descriptor is obtained from the trained networks as follows: first, the extended bounding box of the face is resized so that the shorter side is 256 pixels; then, the center cropping of the face image is carried out after an enlargement of 20% in each direction, in order to include a piece of background. This process made the informative content compliant with the samples used to train the deep neural network. The outcome of the cropping is used as input to the network. The face descriptor is extracted from the layer adjacent to the classifier layer. This leads to a 256 dimensional descriptor. The vector is finally normalized using L2 norm.

4. Experiments and Results

Experiments were carried out on two different datasets. The first one, named **Dataset #1**, contains a collection of videos recorded during debating sessions of the Canary Islands Parliament in Spain. Videos are available at `https://www.parcan.es/video`. These videos contain a large number of shots and different persons involved in a politic debate. In particular, six videos of this huge dataset have been processed. Three of the selected videos have a duration of about 10 minutes (9m43s, 9m41s and 11m23s respectively), whereas the remaining ones are shorter (about 5m each). Videos were recorded on different days and this choice was made in order to test the system on a large number of persons, with different facial appearance (glasses, hairstyle, beard) and to consider different framing conditions and angles. In Figure 3, some frames extracted from **Dataset #1** is reported. As can be observed, videos in this dataset have a number of different camera shots. Each changing in the camera view is totally unrelated to the audio, thus the use of multimedia information would lead the traditional approaches to fail in the identification of the person framed.

The second dataset (**Dataset #2**) contains two videos recorded during therapeutic sessions for assessing the Autism Spectrum Disorders in children at an healthcare centres located in Alessano, province of Lecce, Italy. The above center offers intervention programs for children with Autism Spectrum Disorders diagnosis and/or other disorders. The two videos frame 17 children (14 boys) with Autism Spectrum Disorder diagnosis, aged 6–13 years (Mean = 8.94; Standard Deviation = 2.41) without cognitive delay. The "L'Adelfia" Ethics Committee gave approval for this study and informed signed consent was obtained from

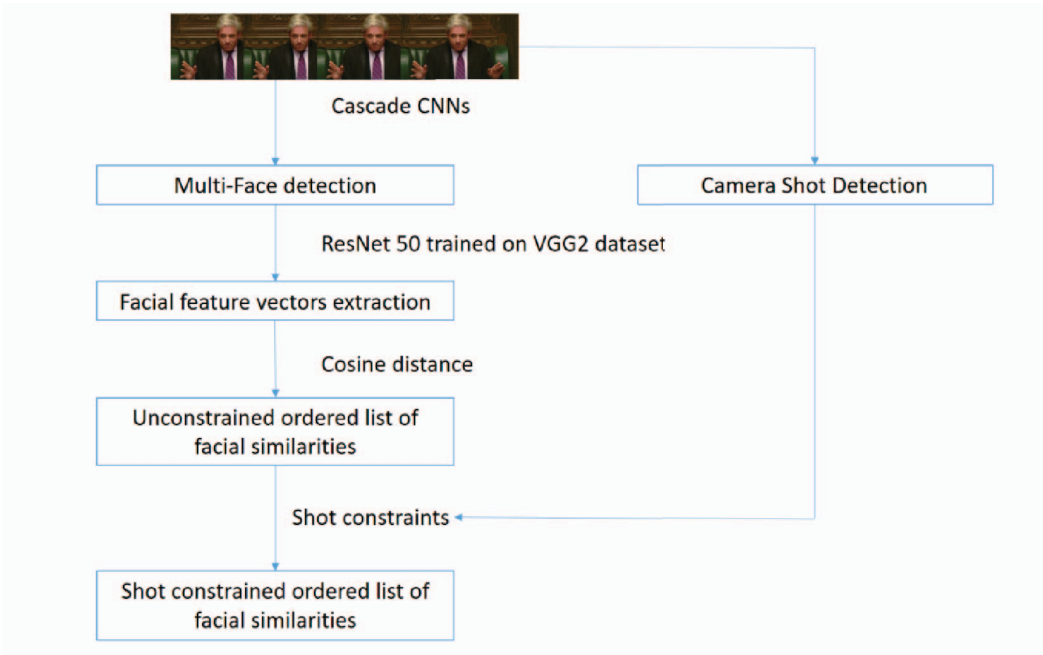

Figure 1. A block diagram of the proposed solution.

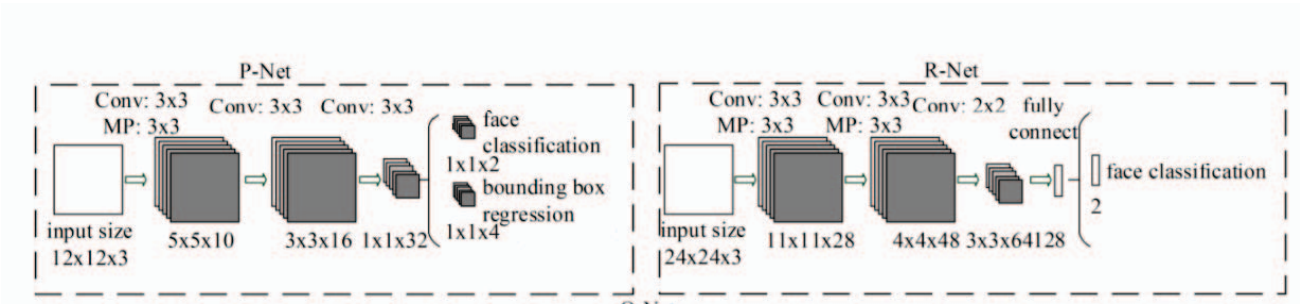

Figure 2. The configuration of the the CNNs exploited in the multi-face detection step.

Figure 3. Some frames extracted from **Dataset #1** composed by Canary Islands Parliament video sessions.

parents. Videos in this second dataset are longer than in the former. Each video lasts more than two hours. In each video, different children alternate in front of a static webcam to perform training sessions aiming at improving their skills in understanding and producing facial expressions. Until automatic tools will be available to analyse acquired data [22], after the session the caregiver could require to retrieve images related to a specific child in order to better assess some aspects of the session but, unfortunately, without automatic indexing, the caregiver needs to scroll the whole video with a large wasting of time and to manually identify the frames where single child appears. Unfortunately, traditional indexing approaches fail in long sequences since, also in this application case, there is no correspondence between audio and person framed (often there is no audio at all) since the therapist's voice is the same for different children. From a technical point of view, in these videos, there is only one camera view (the camera was fixed on a tripod) and one child framed at a time. Children alternated in front of the camera to perform their therapeutic training exercises. The

Table 1. Ground truth data for processed videos. Video identifiers starting with P refer to **Dataset #1** (P stands for Parliament) whereas whose starting with A refer to **Dataset #2** (A stands for Autism).

Video id	Frames	Persons	Camera Shots
P1	8808	8	12
P2	11535	12	18
P3	7412	6	16
P4	6541	7	12
P5	8755	9	11
P6	7996	5	5
A1	217456	4	1
A2	389000	2	1

children do not speak during most of the time, but they try to accomplish the task that is asked of them, that is, to recognize and produce facial expressions. For privacy issues, images of this dataset cannot be published.

Ground truth data for both datasets are reported in Table 1. Table reports, from left to right, the identification label of each video (first column), the number of frames (second column), the number of appearing persons (third column) and the number of different camera shots (fourth column). It is important to underline that a person was considered in the scene only if his face, at least in one frame, was framed with a minimum resolution of 48×48 pixels.

In the experimental phase, each video was fully processed using the pipeline described in Section 3. For videos in **Dataset #2**, the shot transition detector has not been used, manually settings all the frames as belonging to the same shot. The facial patches with and corresponding feature vectors (having 256 items each) were extracted using algorithms explained in Section 3.3, whereas the time bounds of each camera shot were extracted using the methodology described in Section 3.1.

Before carrying out the experimental tests for the whole pipeline for people indexing purposes, a preliminary proof aiming at checking the accuracy of the algorithm implemented for multi-face detection was done. Face detection accuracy was evaluated by annotating 1% of the images in each video and then comparing annotated facial bounding boxes with those automatically labeled by the detector introduced in Section 3.2. In this evaluation step, extracted facial regions that have the Intersection-over-Union (IoU) ratio less than 0.5 to any ground-truth faces were considered as False Positives (FP), whereas extracted facial regions with IoU above 0.5 to a ground truth face were considered as True Positives (TP). As a result of the above, any ground-truth face not having a corresponding detected face was accounted as a False Negative (FN). Table 4 sums up the face detection accuracy according to the aforementioned criteria. From the table it is evident that the face detection worked very well on the considered videos. In particular it

Table 2. Evaluation of the face detector on the considered datasets.

Video id	GT Face	TP	FP	FN	P	R
P1	103	92	15	11	0.85	0.89
P2	155	121	18	34	0.78	0.87
P3	201	174	16	27	0.86	0.91
P4	75	65	12	10	0.86	0.84
P5	121	98	11	23	0.81	0.90
P6	69	61	5	8	0.81	0.92
A1	450	430	1	20	0.95	0.99
A2	500	485	4	15	0.97	0.99
Total	1674	1526	82	148	0.91	0.95

is very accurate on videos $A1$ and $A2$ since they have a simpler background and only one person at time in the scene. Some faces were not detected by the system due to large occlusions or high facial orientation angles. Some examples are reported in Figure 4. Mostly of false positives correspond to incorrect bounding box positioning that brought to IoU value lower than considered threshold.

Figure 4. Some faces that were not detected by the system.

Following this very satisfactory intermediate evaluation, the whole pipeline was then tested. This final evaluation was carried out by computing the cosine distance between feature vectors pertinent to facial patches extracted from the same video. In Figure 5, a graphical representation of the distances computed on the 12670 facial patches detected in the video $P1$ in Table 1 is reported. The figure qualitatively points out how distance values change accordingly to the framed persons. First of all, it is worth noting that blue values follow the diagonal and this means that lowest distance values are properly related to framed persons. Going into details, starting from the top-left corner, the blue (with some shade of light-blue) rectangle represents about the first two minutes of the video in which the President of the Canary Parliament is most of the time framed. The cosine distance is in this case very small since, given the

considered colormap, the blue color is associated to the distance values lower than 0.4. After about two minutes, the parliamentary discussion starts and the camera moves on the 5 different persons involved at different zoom levels (approximately from minute 2 to minute 3.30). Then the President is newly framed for a few seconds, and then a new depute is mainly framed till the end of the video. In Figure 6, the facial patches (sub-sampled with a ratio of 1 every 25 extracted) are plotted along time with Y-axis representing the magnified ($\times 1000$) cosine distance from the first detected face (i.e. the President of the Parliament). This figure clarify even better that the cosine distance on facial feature vectors is very effective to distinguish between the different framed persons.

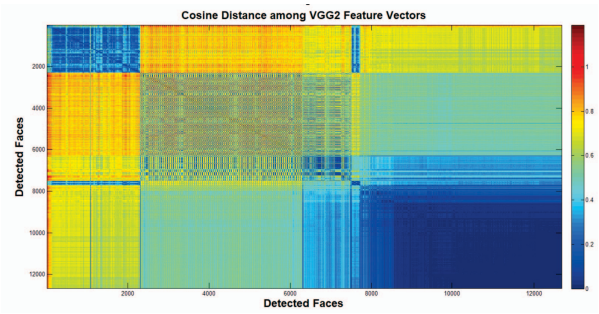

Figure 5. Matrix of the Cosine distances among faces detected in video $P1$ of the Canary Island Parliament.

Concerning the automatic detection of camera shots, it was used only on the parliamentary videos since videos in the second dataset have only one camera shot each. All the camera shots of the videos in **Dataset #1** were correctly detected (100% detection rate).

Finally, for assess the whole framework, the following indexing metrics were considered:

- True re-identification (TR), also known as true match or true positive: a true re-identification occurs when faces that have the same id label (i.e. low cosine distance) belong to the same person;
- True distinction (TD), also known as true non-match or true negative: a true distinction occurs when faces that have different idlabels belong to different persons;
- False re-identification (FR), also known as false positive or false match: a false re-identification occurs when faces with the same idlabels belong to different persons;
- False distinction (FD), also known as false negative or false non-match: a false distinction occurs when faces that have different idlabels belong to the same person.

In this paper the True re-identification (TR) score has been considered and system's outcomes evaluated when using only cosine distance on facial features extracted by ResNet-50 and by integrating information about camera shot transition detection. In the experimental phase, person identification labels were automatically assigned to each persons according to the order of appearance in the video. This means that the first detected face is identified by $id = 1$, the second one has $id = 1$ if its distance from face with $id = 1$ is lower than a given threshold or it has $id = 2$ otherwise, and so on. The i-th face has the label equal to that one of the face having the minimum distance (anyway below the threshold). If there are no close faces in the considered feature space, then a new identifier is introduced. The threshold was set to 0.4 when no information about shot transition detection were used. Two adaptive thresholds were instead introduced to get advantage of the knowledge of the camera shot boundaries. The aforementioned threshold was in fact relaxed to 0.6 when the comparison was made between two face images belonging to the same shot.

At the end, assigned label were compared with true labels manually annotated and the results of this comparison is reported in Figure 7. The advantages in using shot transition detection are evident. Many errors in face matching are avoided by considering different thresholds depending on the availability of shot transition detector information.

About implementation details, the face detector has been implemented in C++. A modified version of the pre-trained CAFFE model of [36] has been used. For facial vector generation, a SE-ResNet-50-256D has been implemented in CAFFE using the pre-trained models built on VGGFace2 dataset as introduced in [5]. The camera shot transition detection has been written in C++ with the OpenCV libraries, while the face indexing illustrated in the last two blocks has been implemented with MATLAB 2019a.

5. Conclusion

Most relevant works in the literature gather full indexing of videos in real scenarios by exploiting additional media features (e.g. audio and text), but there are application contexts where multimedia data are either not available or reliable. At this aim, a new fully computer vision based video indexing pipeline has been proposed in this work. The proposed pipeline can detect and re-identify people in each image supplying an automatic subject dependent video indexing. Faces are processed without any constraints about appearance, eyeglasses, beard or hairstyle. Moreover, a shot transition detector has been introduced and employed to strengthen the results. The system has been validated and tested in two different typical scenarios where no-multimedia data could be exploited: broadcasted political video documentaries and healthcare therapies sessions about non-verbal skills. Experiments demonstrate encouraging results in the field of on-demand video retrieval, making the system suitable in the two assistive scenarios under

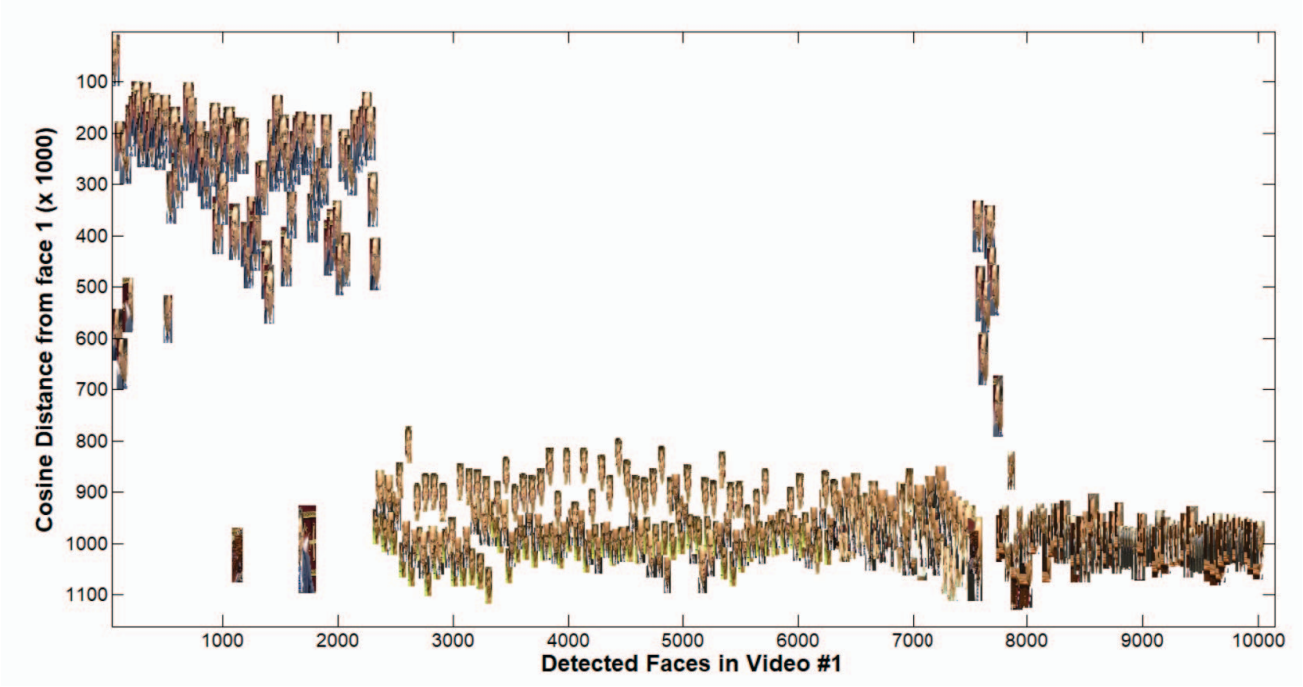

Figure 6. Graphical representation of faces in the first video of **Dataset #1** (P1) with respect to the cosine distance from the first detected face.

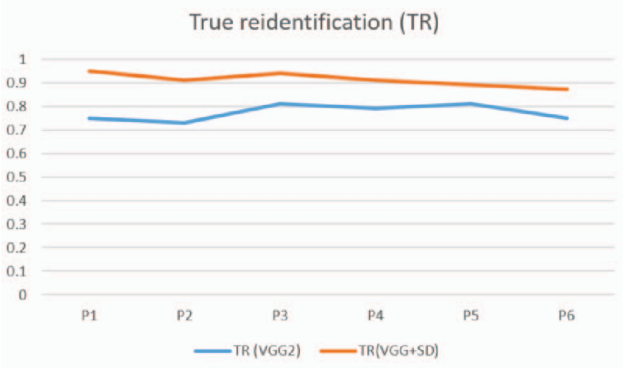

Figure 7. System's performance in indexing of persons with and without taking into account camera shot boundaries.

investigation. Future works will consider the possibility to put in place the proposed approach as part of an assistive system to be exploited in healthcare centers for assessing and diagnosing Autism Spectrum Disorders in children.

References

[1] S. Abriola, P. Barceló, D. Figueira, and S. Figueira. Bisimulations on data graphs. *Journal of Artificial Intelligence Research*, 61:171–213, 2018.

[2] S. Bagheri, J. Y. Zheng, and S. Sinha. Temporal mapping of surveillance video for indexing and summarization. *Computer Vision and Image Understanding*, 144:237–257, 2016.

[3] K. Bougiatiotis and T. Giannakopoulos. Enhanced movie content similarity based on textual, auditory and visual information. *Expert Systems with Applications*, 96:86–102, 2018.

[4] H. Bredin and G. Gelly. Improving speaker diarization of tv series using talking-face detection and clustering. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 157–161. ACM, 2016.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[6] D. Cazzato, M. Leo, and C. Distante. A complete framework for fully-automatic people indexing in generic videos. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 248–255. IEEE, 2014.

[7] J. Y. Choi, K. N. Plataniotis, and Y. M. Ro. Face annotation for online personal videos using color feature fusion based face recognition. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1190–1195. IEEE, 2010.

[8] F. Chowdhury, Q. Wang, I. L. Moreno, and L. Wan. Attention-based models for text-dependent speaker verification. *arXiv preprint arXiv:1710.10470*, 2017.

[9] E. Corvee, F. Bremond, M. Thonnat, et al. Person re-identification using haar-based and dcd-based signature. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–8. IEEE, 2010.

[10] Z. Dong, C. Jing, M. Pei, and Y. Jia. Deep cnn based binary hash video representations for face retrieval. *Pattern Recog-*

nition, 81:357–369, 2018.

[11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010.

[12] G. Friedland, H. Hung, and C. Yeo. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4069–4072. IEEE, 2009.

[13] V. Gandhi and R. Ronfard. Detecting and naming actors in movies using generative appearance models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3706–3713, 2013.

[14] I. D. Gebru, S. Ba, X. Li, and R. Horaud. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *arXiv preprint arXiv:1603.09725*, 2016.

[15] I. D. Gebru, S. Ba, X. Li, and R. Horaud. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1086–1099, 2018.

[16] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1528–1535. IEEE, 2006.

[17] I. U. Haq, K. Muhammad, A. Ullah, and S. W. Baik. Deepstar: Detecting starring characters in movies. *IEEE Access*, 7:9265–9272, 2019.

[18] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017.

[19] T. Huang and S. Russell. Object identification in a bayesian context. In *IJCAI*, volume 97, pages 1276–1282, 1997.

[20] K. Kim, Z. Yang, I. Masi, R. Nevatia, and G. Medioni. Face and body association for video-based face recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 39–48. IEEE, 2018.

[21] P. Kulkarni, B. Patil, and B. Joglekar. An effective content based video analysis and retrieval using pattern indexing techniques. In *Industrial Instrumentation and Control (ICIC), 2015 International Conference on*, pages 87–92. IEEE, 2015.

[22] M. Leo, P. Carcagnì, C. Distante, P. Spagnolo, P. Mazzeo, A. Rosato, S. Petrocchi, C. Pellegrino, A. Levante, F. De Lumè, et al. Computational assessment of facial expression production in asd children. *Sensors*, 18(11):3993, 2018.

[23] P. Li, J. Xie, Z. Li, T. Liu, and W. Yan. Facial peculiarity retrieval via deep neural networks fusion. *International Journal of Computational Intelligence Systems*, 11(1):58–65, 2018.

[24] X. Liu, J. Geng, H. Ling, and Y. ming Cheung. Attention guided deep audio-face fusion for efficient speaker naming. *Pattern Recognition*, 88:557 – 568, 2019.

[25] N. E. Maliki, H. Silkan, and M. E. Maghri. Efficient indexing and similarity search using the geometric near-neighbor access tree (gnat) for face-images data. *Procedia Computer Science*, 148:600 – 609, 2019. THE SECOND INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING IN DATA SCIENCES, ICDS2018.

[26] T. D. Ngo, H. T. Vu, D.-D. Le, and S. Satoh. Face retrieval in large-scale news video datasets. *IEICE TRANSACTIONS on Information and Systems*, 96(8):1811–1825, 2013.

[27] S. Pini, M. Cornia, F. Bolelli, L. Baraldi, and R. Cucchiara. M-vad names: a dataset for video captioning with naming. *Multimedia Tools and Applications*, Dec 2018.

[28] J. Prinosil. Blind face indexing in video. In *Telecommunications and Signal Processing (TSP), 2011 34th International Conference on*, pages 575–578. IEEE, 2011.

[29] M. Ravinder and T. Venugopal. Content-based video indexing and retrieval using key frames texture, edge and motion features. 2016.

[30] E. Sánchez-Nielsen, F. Chávez-Gutiérrez, J. Lorenzo-Navarro, and M. Castrillón-Santana. A multimedia system to produce and deliver video fragments on demand on parliamentary websites. *Multimedia Tools and Applications*, 76(5):6281–6307, Mar 2017.

[31] N. Sarafianos, T. Giannakopoulos, and S. Petridis. Audio-visual speaker diarization using fisher linear semi-discriminant analysis. *Multimedia Tools and Applications*, 75(1):115–130, 2016.

[32] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping. In *European Conference on Computer Vision*, pages 423–432. Springer, 2012.

[33] J. Tang, Z. Li, and X. Zhu. Supervised deep hashing for scalable face image retrieval. *Pattern Recognition*, 75:25–32, 2018.

[34] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014.

[35] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1353, 2016.

[36] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[37] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[38] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.