Spatio-temporal Localization of Actors in Video/360-Video and its Applications

Paulo Mendes
paulo.mendes@telemidia.puc-rio.br
TeleMídia/PUC-Rio
Rio de Janeiro, Brazil

Sérgio Colcher
colcher@inf.puc-rio.br
Informatics Departament/PUC-Rio
Rio de Janeiro, Brazil

Abstract

The popularity of platforms for storing and transmitting video content has created a substantial volume of video data. Given a set of actors present in a video, generating metadata with the temporal determination of the interval in which each actor is present and their spatial 2D localization in each frame in these intervals can facilitate video retrieval and recommendation. In this work, we investigate *Video Face Clustering* for this spatio-temporal localization of actors in videos. We first describe our method for *Video Face Clustering* in which we take advantage of face detection, embeddings, and clustering methods to group similar faces of actors in different frames and provide the spatio-temporal localization of them. Then, we explore, propose, and investigate innovative applications of this spatio-temporal localization in three different tasks: (i) *Video Face Recognition*, (ii) *Educational Video Recommendation* and (iii) *Subtitles Positioning in 360-video*.

Keywords: clustering, face recognition, video recommendation, 360-video, multimedia authoring

1 Introduction

In recent years, the popularity of platforms for storing and transmitting video content has stimulated the production of a massive volume of video data, establishing new habits and leveraging new applications with innovative forms of consumption of this kind of information. As an indication of this huge production (and consumption) of data, in 2019, more than one billion hours of YouTube videos were watched per day.[1]

Generating metadata with (i) the identity information of actors present, (ii) the temporal determination of the intervals in which each of these actors is present, and (iii) their spatial localization in each of the frames along these intervals can facilitate video indexing, retrieval, recommendation and a series of other tasks which might enhance the way people interact and consume all this video data. Besides the identification, (ii) and (iii) together are what we call *Spatio-temporal Localization*.

In this dissertation, we investigate a method for the spatio-temporal localization of actors in videos. Our expected contribution is two-fold: (1) we propose a core process for the spatio-temporal localization in which we take advantage of face detection, embeddings, and clustering methods to group similar faces (presumably from the same actors) along with the frames, and (2) we further propose and explore the innovative application of this localization in three different practical and important tasks: *Video Face Recognition*, *Educational Video Recommendation*, and *Subtitles Positioning in 360-video*.

2 A Method For Video Face Clustering

This section describes the core of this dissertation, which is a method for *Video Face Clustering*. It consists of detecting and grouping faces from different video frames (ideally from the same actors) extracted from a video file. Figure 1 depicts this process, and each of its steps is described in the following paragraphs.

First, we perform *Frames Extraction* by receiving a video file as input and extracting its frames according to a given frame rate. These frames are used as a set of images for the next step.

The *Face Detection* step uses an object detection model for detecting faces in each of its images. In our case, objects are faces and, therefore, the face detection model is responsible for returning the bounding boxes of the faces present in the image, specified by the x and y axes coordinates of the upper-left corner and of the lower-right corner of the rectangle that establishes the visual limits that encapsulate each face. With these bounding boxes, we can isolate and extract the bounded images, obtaining a dataset composed of images of faces only.

The objective of the *Embeddings Generation* step is to represent each face image as a vector in $\mathbb{R}^n$. To achieve that, it processes each of the faces generated in the previous step through a CNN, generating embeddings. Ideally, an embedding captures some semantics of the input, e.g., by placing semantically similar inputs close together in the embedding space. At the end of this step, we have all faces represented as embeddings.

[1] https://kinsta.com/blog/youtube-stats/

In: IV Concurso de Teses e Dissertações (CTD 2022), Curitiba, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2022.

ISSN 2596-1683

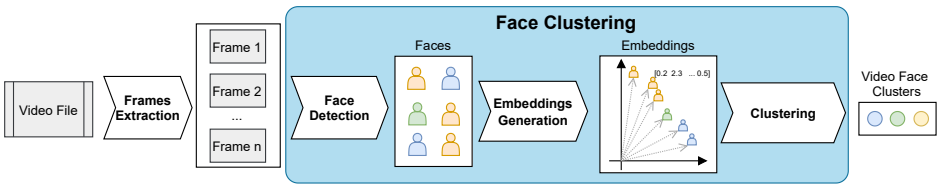


Figure 1. Video face clustering process.

In the *Clustering* step, we group embeddings and, consequently, faces that are close in the embedding space using a clustering algorithm.

The clustering process should produce a partition of the dataset, i.e., each generated cluster represents a specific person, and the union of all clusters covers the whole dataset.

The following three sections describe the applications we investigate in this dissertation. These three applications propose novel approaches for tasks in video using spatio-temporal localization of actors through *Video Face Clustering*.

3 Cluster-Matching-Based Method For Video Face Recognition

This section describes the first application we investigated using *Video Face Clustering*. We propose a cluster-matching-based approach for video face recognition where *face clustering* is used to group faces in both the face dataset and in the target video (*video face clustering*). Consequently, classes do not have to be previously known, and the effort spent with annotations is significantly reduced — as it is done over clusters instead of single images. Face recognition becomes a task of comparing clusters from the dataset with those extracted from images or video sources. Therefore, our approach is easily scalable. Our method intends to recognize people in video using *Video Face Clustering* and a matching heuristic. Figure 2 shows our proposed approach.

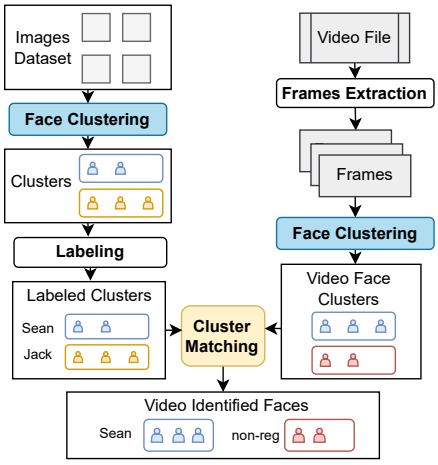


Figure 2. Cluster-Matching based Method for Video Face Recognition.

In our approach, we use *Face Clustering* in the images dataset and the referenced video. Then, the clusters of the images dataset are labeled. In the *Labeling* step, we assign labels (identities) to represent the clusters. Using this pipeline, instead of having to label every single face for constructing a labeled dataset, it is only necessary to label each generated cluster. Hence, the labeling complexity becomes a function dependent on the number of clusters, which is at most as great as the number of individuals. At the end of this step, we have a dataset of *Labeled Clusters*.

Next, the *Cluster Matching* step receives the set of clusters from the video and the set of labeled clusters, which is used as a reference for recognizing the clusters (and consequently the faces) in the video. We designed a method based on cluster distance for performing this recognition using a representative embedding for each cluster.

Our method can be used to generate metadata in video files indicating the people that appear on them. Figure 3 shows the first 12 seconds of a video where two identified Brazilian politicians are shown in each frame with their respective colored clusters. The results of this application were published at a relevant multimedia conference [2].

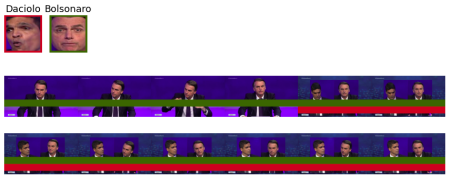


Figure 3. Timeline with tagged frames by their clusters of registered people

4 Face-Clustering-Based Method for Educational Video Recommendation

This section describes the second application we investigated using our *Video Face Clustering* method. We propose a recommender method based on the actor's presence for educational videos. In this case, the actors are lecturers (or teachers, professors, etc.) presenting educational content on video. For instance, if a student watches a video containing lecturers A and B, our method recommends other videos containing at least one of these lecturers. This method provides an additional aid for educational recommender systems, allowing them to use the presence of lecturers as a feature for composing their recommendations.

We first represent each video in the dataset of educational videos by the clusters of lecturers present using *Video Face Clustering*. Once we have all videos represented, we perform the pipeline depicted in Figure 4.

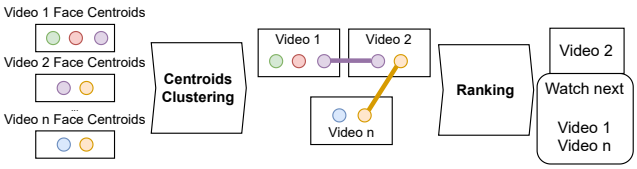


Figure 4. Video Recommendation based on Lecturers Centroids Clustering

First, we gather the centroids from the videos of the dataset as one single set and perform the *Centroids Clustering*.

In doing that, we group centroids from the same lecturer from different videos. For instance, in Figure 4, one can see that the *purple lecturer* is present in both Videos 1 and 2, while the *orange lecturer* is present in both Videos 2 and n.

Next, based on these relationships among different videos, we perform *Ranking* by recommending videos in which lecturers of the current video are present. For doing that, we compute a similarity score using the presence of the lecturers in the current video and the presence of these same lecturers in the other video so that the more lecturers they have in common and the more they appear, the greatest this score will be.

Finally, using this score, we compute a ranking for each video with the greatest scores as the recommended videos. By the end of this phase, we have a ranking of recommended videos for each video in the dataset. It is important to notice that our method is unsupervised and does not require the information of the lecturers in advance. We have also published these results in a relevant multimedia conference [4].

5 Subtitles Positioning in 360-Videos

In [3], we proposed an authoring model for interactive 360-video. In such a model, we can define interactive 360-videos that are presented together with additional information, such as images, text, subtitles, 2D traditional videos, and spatial audio. We use polar coordinates to define the positioning of such information, along with a start time and a duration. For instance, we can define that a text moves with the user's head motion and is always visible or that such text is placed at a fixed position if it is in the user's field of view. In this section, we describe how *video face clustering* can be used with this authoring model for automatic subtitles positioning in 360-video.

Nowadays, the most common way of representing and transmitting 360-video is using equirectangular projection [5]. This kind of projection creates challenges for image processing and computer vision algorithms, especially convolution-based ones, because of the severe distortions in areas vertically distant from the center of the image (see Figure 5). Due to these distortions, we adapted the *Face Detection* step of our method, which uses a traditional CNN.

Figure 5. Example of 360-image represented through equirectangular projection.

In order to mitigate the problem of severe distortions present in the equirectangular projection, we have used an approach based on viewports extraction. Figure 6 shows this process. Given an equirectangular image, which could also be a frame from a 360-video, we first perform *Viewports Extraction*. A viewport represents a portion of the 360-degree scene, similar to when a picture is taken in the real world. The picture represents the world from the point of view of a specific direction. With these viewports, we aim to reduce the distortions since we are representing the equirectangular image with a series of standard images, similar to the ones used by traditional CNNs. Next, we perform *Face Detection* using a traditional CNN for each of these viewports. Hence, for each viewport, we have a group of faces detected.

In the last step of our approach, called *Mapping*, we map each face detected back to the equirectangular image.

Using these adaptations, we can apply *video face clustering* in 360-videos. With this clustering, for each actor detected, we generate a file containing his/her position (in polar coordinates) for each point of time.. By doing so, we use these files for defining the subtitles positioning using our authoring model. We have developed a player based on the Unity Game Engine[2] to support this model. We can also define a more interactive setting in which the subtitles follow the actors in the user's viewport, which means that the user can see them. When they are not visible, subtitles are presented at the bottom center of the user's viewport. Figure 7 shows how this setting works. In Figure 7a, the subtitles are placed close to the actor visible in the current user's viewport. When the actor is not visible, as shown in Figure 7b, the subtitles are placed at the bottom of the user's viewport. In this way, the user does not lose the content present in the subtitles. A video demonstration can also be seen on YouTube.[3]

6 Discussion

In this dissertation, we present a method for spatio-temporal localization of actors based on *Video Face Clustering*. As part of this method, we also define an algorithm for finding an adequate number of clusters based on the silhouette score. We investigate to what extent this method can be used as the

[2]https://github.com/TeleMidia/VR360Authoring
[3]https://youtu.be/lJKhgZl1Rc8

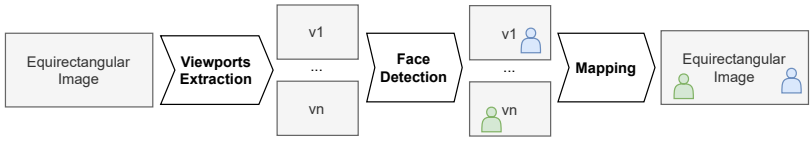
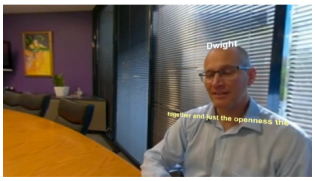
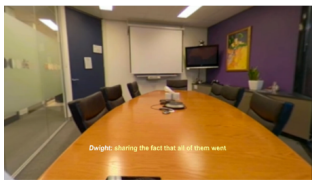


Figure 6. 360-degree face detection process.



(a) Subtitles close to actors' faces when they are in the user's viewport.



(b) Subtitles positioning in the bottom of user's viewport otherwise.

Figure 7. Dynamic subtitles positioning using video face clustering and our authoring model.

core to leverage and enhance some innovative applications, especially in three different practical and important tasks: *Video Face Recognition* [2], *Educational Video Recommendation* [4], and *Subtitles Positioning in 360-Video* [3].

For the *Video Face Recognition*, we propose a cluster-matching-based approach, derived mainly from the characteristics of our core *Video Face Clustering* method, which is very scalable since the effort spent with annotations is significantly reduced — as it is done over clusters instead of single images. This method uses *Video Face Clustering* and a heuristic for cluster matching to recognize people in videos. It has achieved a recall of 99.435% and precision of 99.131% when considering faces extracted from a set of 13 video files. As another consequence of face clustering, our technique can be helpful for creating and labeling datasets in a less time-consuming way by labeling clusters instead of individual images.

For the *Educational Video Recommendation* task, we investigate a new feature: the presence of specific lecturers, which again is a direct result of the application of our core method. After performing *Video Face Clustering* on each video, we extract their centroids to perform another clustering step that creates a relationship of videos that share the presence of the same lecturers. Finally, we rank the recommended videos based on the amount of time that each lecturer is present. Our method uses only the video files for performing recommendations; no other information about these videos nor the identity of the lecturers is necessary. It is worth mentioning that we do not intend to substitute other video recommendation methods. Instead, our application shows that if the presence of lecturers is a relevant feature for educational video recommendation, it can be used for this purpose with a mAP value of 0.99.

For the *Subtitles Positioning in 360-Video* task, our main contribution is the proposal of dynamic placement of subtitles based on the automatic localization of actors. To achieve this goal, we adapted our spatio-temporal localization method to the 360-video setting and created an authoring model for interactive 360-videos. Because of the severe distortions in equirectangular 360-videos, we used an approach based on viewports extraction for the face detection step of *Video Face Clustering*. To evaluate this approach against the sole use of a traditional CNN, we created a synthetic dataset by projecting images from the FDDB benchmark [1] to equirectangular backgrounds.

Moreover, we proposed an authoring model that allows authors to design and create interactive 360 videos. The proposed model is the result of the analysis of different immersive 360 multimedia applications scenarios. Besides supporting subtitles positioning in 360-videos, it also supports navigation among 360-videos, additional media, etc. In summary, we highlight the following contributions of this dissertation: (1) a method for video face recognition that is easily scalable and helps in labeling images dataset; (2) a method for educational video recommendation based on the presence of lecturers; (3) method for face detection in equirectangular 360-images that uses models pre-trained with traditional images; (4) a synthetic dataset for face detection in equirectangular 360-images; (5) an authoring model for interactive 360-videos; (6) a player for interactive 360-videos; (7) an approach for automatic positioning subtitles in 360-videos based on the actors' positions;

Acknowledgments

This work was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil.

References

[1] Vidit Jain and Erik Learned-Miller. 2010. *Fddb: A benchmark for face detection in unconstrained settings.* Technical Report. UMass Amherst technical report.

[2] P. Mendes, A. Busson, S. Colcher, D. Schwabe, A. Guedes, and C. Laufer. 2020. A Cluster-Matching-Based Method for Video Face Recognition. In *Proceedings of the Brazilian Symposium on Multimedia and the Web.* 97–104.

[3] P. Mendes, A. Guedes, D. Moraes, R. Azevedo, and S. Colcher. 2020. An Authoring Model for Interactive 360 Videos. In *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW).* 1–6. https://doi.org/10.1109/ICMEW46912.2020.9105958

[4] P. Mendes, E. Vieira, A Guedes, A. Busson, and S. Colcher. 2020. A Clustering-Based Method for Automatic Educational Video Recommendation Using Deep Face-Features of Lecturers. In *2020 IEEE International Symposium on Multimedia (ISM).* 158–161. https://doi.org/10.1109/ISM.2020.00034

[5] Wenyan Yang, Yanlin Qian, Joni-Kristian Kämäräinen, Francesco Cricri, and Lixin Fan. 2018. Object detection in equirectangular panorama. In *2018 24th International Conference on Pattern Recognition (ICPR).* IEEE, 2190–2195.