Image and Vision Computing 57 (2017) 25–43

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Exploiting scene maps and spatial relationships in quasi-static scenes for video face clustering☆

Alessio Bazzica*, Cynthia C.S. Liem, Alan Hanjalic

Delft University of Technology, Multimedia Computing Group, Mekelweg 4, Delft 2628 CD, The Netherlands

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form 14 September 2016
Accepted 8 November 2016
Available online 23 November 2016

Keywords:
Video face annotation
Face clustering
Re-identification

ABSTRACT

Video face clustering is a fundamental step in automatically annotating a video in terms of when and where (i.e., in which video shot and where in a video frame) a given person is visible. State-of-the-art face clustering solutions typically rely on the information derived from visual appearances of the face images. This is challenging because of a high degree of variation in these visual appearances due to factors like scale, viewpoint, head pose and facial expression. As a result, either the generated face clusters are not sufficiently pure, or their number is much higher than that of people appearing in the video. A possible way towards improved clustering performance is to analyze visual appearances of faces in specific contexts and take the contextual information into account when designing the clustering algorithm. In this paper, we focus on the context of *quasi-static scenes*, in which we can assume that the people's positions in a scene are (quasi-)stationary. We present a novel video clustering algorithm that exploits this property to match faces and efficiently propagate face labels across the scope of viewpoints, scale and level of zoom characterizing different frames and shots of a video. We also present a novel publicly available dataset of manually annotated quasi-static scene videos. Experimental assessment on the latter indicates that exploiting information derived by the scene and the spatial relationships between people can substantially improve the clustering performance compared to the state-of-the-art in the field.

1. Introduction

Annotating where and when a person appears in a video over time can be beneficial in many applications including video retrieval [11] and video surveillance [19]. This type of annotation essentially involves two steps, namely detecting a person in a video frame and linking the appearances of the same person across different frames. These operations can be done manually, automatically, or semi-automatically.

Manual annotation can be accurate since a human annotator can easily spot faces in the wild and say whether two faces belong to the same person. However, a fully manual annotation process is typically tedious and time consuming. As an alternative, automatic methods can be applied; they typically work by first detecting people via face detection and tracking, and then by linking the detected faces either by recognition [35] or clustering [8,11,23,28]. Fully automatic methods scale well with long videos, but they may not guarantee a satisfying performance as several types of errors can occur. For instance, (profile) faces may not be detected, leading to lack of *coverage*[25]. Non-face regions can be mistaken for faces and face trackers can drift. Face recognition may also fail, especially when the face images are not detailed enough, and it can only work for known people, for which identity models need to be built using a collection of labeled images. Face clustering solutions are more generally applicable as they do not rely on trained identity models, but they typically suffer from the sub-clustering problem, which occurs when different visual appearances of one identity are recognized as different people.

Given the current caveats of fully manual and automatic methods, the most viable approach for person annotation in practice may lie somewhere in-between, in the form of semi-automatic methods that effectively combine automatic computation and human annotation [5,33]. Such methods can be further optimized by improving, in particular, the reliability of automatic clustering modules in order to significantly reduce the human effort.

In view of the above, automatic face clustering methods are critical for either fully automated or semi-automated face annotation scenarios and there are still numerous challenges to be pursued in order to bring the robustness and reliability of such methods to an acceptable level. In this paper we provide a novel contribution in this direction.

☆ This paper has been recommended for acceptance by Xiaogang Wang.

* Corresponding author.
E-mail addresses: A.Bazzica@tudelft.nl (A. Bazzica), C.C.S.Liem@tudelft.nl (C. Liem), A.Hanjalic@tudelft.nl (A. Hanjalic).

http://dx.doi.org/10.1016/j.imavis.2016.11.005
0262-8856/

A logical starting point to develop an automatic face clustering method is the information derived from the visual appearance of a face in a video frame. Relying on this source alone has, however, been shown to be unreliable due to a large degree of variations of faces' visual appearances across video frames and shots. This is due to the factors like camera viewpoint, face pose (e.g., profile vs. frontal face) and face scale (e.g., close-up vs. group of people) [32]. To cope with these challenges, contextual information could be exploited that characterizes specific categories of video.

In [10], visual features related to clothing proved to be beneficial in the case of TV talk shows. Furthermore, when the audio channel is closely coupled to the visual one, as in the case of a debate with the active speaker being filmed, speech features could strengthen the link between two people's appearances if the visual channel is not sufficiently informative [11]. While these approaches may lead to better performance, they are not applicable to all videos since the required information may not always be available. Methods relying on visual context information (e.g., clothing) may still suffer from the challenges of matching images at multiple scales and from multiple camera angles. Besides, there are cases in which the additional analyzed cues may not be discriminative enough to distinguish different subjects (e.g., similar clothing, similar voice).

Another step towards a more comprehensive solution is including contextual information which still is available in the visual domain, but related to the environment. For instance, people co-occurrence patterns [30,32] can be used to link face sub-clusters. This approach can help to improve the recall. However, it relies on the assumption that the sub-clusters to merge are nearly 100% pure, which is difficult to guarantee in practice [6].

In this work, we propose a video face clustering method that exploits contextual information derived from the relative position of people appearing in *quasi-static scenes* — i.e., scenes in which the spatial configuration of the objects appearing in a video is (quasi-) stationary over a particular time interval. As explained in more detail in Section 3.1, the term "quasi" is used to allow limited variations in the visual appearance of the objects. For instance, the visual appearance of a face may change due to head movement or occlusion and the face object can also slightly move. As shown in Fig. 1, this scene category applies to a broad range of video genres including, for instance, talk shows and TV debates, TV game shows and symphonic concerts.

In addition to being applicable to a large variety of videos, our proposed solution also removes the need for often tedious and time consuming processes of building person identity models. Therefore, although some quasi-static videos (e.g., talk shows and TV debates) may involve well-known people, for whom facial recognition could be applied, we still pursue a generic clustering solution for this entire category.

Finally, our method can handle complex situations, in which the number of people to be annotated is large. For instance, in an overview shot including dozens of people, the visual detail of each face can be quite poor. However, the relative position between two people can still provide useful information to infer the correct identities.

As shown by the example in Fig. 2, the proposed method consists of the following steps:

- a **map of the scene** is learned by finding a set of geometrically consistent matches between keyframes (see views 1, 2, and 3);
- the **regions of visual overlap** between matching keyframes are computed (see the black, blue, and red regions and the arrows connecting them);
- the overlapping keyframe pairs are used to efficiently **match faces** via sub-graph matching [36] (see the white arrows connecting faces across keyframes);
- a **face matches graph** is built and used to derive the final face clusters via connected component analysis.

Following this procedure, our algorithm avoids a brute-force comparison between each face pair and exploits the spatial configuration of the filmed people as information for matching their appearances, making it effective when dealing with different viewpoints, different zoom levels and/or varying head poses and expressions.

To the best of our knowledge, this is the first time that the spatial configuration of the filmed people is exploited as context to enable clustering of faces. To encourage further research, we release the code and a fully annotated dataset, referred to as the *QSS dataset*, which has been built using four YouTube videos and a professional symphonic orchestra video.

The paper is organized as follows. We start by giving an overview of the related work in video face clustering in Section 2. Then, in Section 3, we explain how we address the limitation of the existing approaches and present our method in Section 4. The details about the datasets and the evaluation approach are reported in Section 5, while in Section 6, we assess our method by comparing it to a number of baselines and existing approaches. The conclusions drawn from this comparison are reported in Section 7, together with an outlook towards future research.

2. Related work

Video face clustering can be performed in several ways, depending on the type of information a method exploits. A standard pipeline involves the following steps: (key)frame based face detection, face tracking (typically informed by shot boundaries), visual features extraction and, finally, building face track clusters using similarity scores between faces (or face tracks). A popular paper describing a video face clustering system is [25].

The main distinguishing factor of a method is the type of information and features being exploited. A method can rely on clustering constraints that are automatically generated, and different strategies to compute visual face similarity. Depending on the type of video, contextual information can also be exploited. As anticipated in the example of Fig. 1, our proposed method also relies on two

Fig. 1. Example of quasi-static scenes: the filmed people's spatial configuration is stationary over a particular time interval.