Neurocomputing 152 (2015) 316–321

Contents lists available at ScienceDirect

Neurocomputing

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

Automatic face annotation in TV series by video/script alignment

Yifan Zhang [a], Zhiqiang Tang [a], Chunjie Zhang [b], Jing Liu [a,*], Hanqing Lu [a]

[a] National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 4 September 2014
Received in revised form
10 October 2014
Accepted 23 October 2014
Communicated by X. Li
Available online 13 November 2014

Keywords:
Face annotation
Face naming
Video analysis
Sequence alignment

ABSTRACT

This paper describes a method for automatically tagging the names to the faces which are collected from uncontrolled TV series videos. The detected faces are firstly partitioned into several clusters. Then we construct a face sequence based on their occurrence order in the video and denote them by cluster labels. It can be assumed that the temporal distribution of the faces in the video roughly follows the temporal distribution of the names in the script. Hence, we propose to annotate the faces by video/script alignment. A global sequence alignment algorithm is employed to find the most probable faces in the face sequence matching to the names in the name sequence. The novelty lies in that we consider the temporal order relationship of the faces and names over the whole video and directly align two heterogeneous sequences. Experiments on real-world videos have demonstrated the effectiveness and efficiency of the proposed method.

1. Introduction

This paper investigates the problem of automatically annotating the faces with the character names in TV series videos. The objective is to find all the faces of certain person and to attach the correct name to them. Based on our work, one can easily use the character name as a query to select the characters of interest and view the related video clips. This character-centric browsing is able to provide a new way for video summarization and digestion, thus bringing us a new viewing experience.

The face annotation task is challenging due to high variability of the faces in pose, scale, facial expression, as well as illumination, occlusion and camera movement. In addition, we also have to face the weakness and ambiguity of the available textual information. Since the text is not well temporally aligned with the video, the crux is how to exploit the relationship between the video and the associated text to build the links between the faces and the names.

The previous efforts on video face annotation can be roughly divided into two categories: classification based and clustering based. Most of the previous works focus on person classification. Everingham et al. [1] combine facial and clothing features to train a Bayesian classifier. Sivic et al. [2] use the facial feature to train a multiple kernel SVM classifier. Tapaswi et al. [3] model each TV series episode as a Markov Random Field, integrating face, clothing and speaking features together to increase the coverage of the labeled persons. Bäuml et al. [4] employ multinomial logistic regression classifiers for multi-class classification which incorporates labeled and unlabeled data. Albeit the classification methods can achieve good performance, they typically rely on the quality of the annotated training exemplars for each person. The training data is obtained either by manual labeling or by textual alignment between subtitles and scripts [1]. Everingham et al. [1] align the subtitles with the scripts by dynamic time warping to get the time tags of the names, and then link them to the faces in the videos to get labeled face exemplars. However, the name cues from the textual alignment are weak and ambiguous since faces of speakers may not be visible, or there may be more than one face visible at the same time. In addition, its application scope is limited by the availability of the subtitle, which usually does not exist in many non-European language TV shows [5].

To extend the application scope to the scenarios when subtitles are not available, some researchers investigate the problem on how to build the linking between the faces and the names without using time tags. These works are mainly based on face clustering. Some sophisticated face clustering methods are proposed. Wu et al. [6] use a Markov random field to model the relationships between faces and incorporate pairwise constraints. Lu and Ip [7] propose a constrained spectral clustering. Both of the methods can propagate the constraints to neighboring data based on smooth assumption. Cinbis et al. [8] propose an unsupervised metric learning algorithm for face identification in TV video. All these methods focus on face clustering, they cannot automatically link the faces to their real names. To this end, in [9], we propose a global face–name matching framework, in which the weighted face graph from the video is matched with the weighted name

[*] Corresponding author.
E-mail addresses: yfzhang@nlpr.ia.ac.cn (Y. Zhang), zqtang2013@gmail.com (Z. Tang), cjzhang@jdl.ac.cn (C. Zhang), jliu@nlpr.ia.ac.cn (J. Liu), luhq@nlpr.ia.ac.cn (H. Lu).

http://dx.doi.org/10.1016/j.neucom.2014.10.062
0925-2312/© 2014 Elsevier B.V. All rights reserved.

graph from the script. The weighted face graph is constructed by face clustering. Sang and Xu [10] extend the work of [9] by using an ordinal graph and employing a new graph matching algorithm called Error Correcting Graph Matching. Liang et al. [11] propose a generative model in which character histogram is used to depict the correspondence between the video and the script. The face–name association matrix is automatically learned as the parameters of the model. Generally, most of these methods match the faces and names based on local face–name relationship within each scene of a TV series video or a movie. The global temporal order relationship over the whole video is not used. The method in [11] uses forward traversing and backward tracing algorithms to conduct the matching over the whole video. However, it employs the character histogram to compare the faces and names in shots and scenes. Consequently, most of their performance relies on the scene segmentation results.

In this paper, we present a novel framework (see Fig. 1) for face annotation based on face clustering and global video/script alignment. In the framework, global temporal order relationship over the whole video is fully exploited, without relying on scene segmentation. We firstly build two sequences from the video and the script respectively: a face sequence and a name sequence. In the face sequence, faces are denoted by their cluster labels which are obtained by a clustering algorithm. It can be assumed that the temporal distribution of faces of one character is approximately similar to the temporal distribution of their names along the time line. Hence, the face naming problem can be transformed to a sequence alignment problem. In our previous work [12], we use the Levenshtein distance [13] measurement to find the optimal alignment between the face track sequence and the name sequence. Since the Levenshtein distance is the minimum editing distance between two homogeneous sequences, to extend it to two heterogeneous sequences, we have to enumerate all possible matching results between the face clusters and the names. It quickly becomes computationally intractable when the number of the different elements in the two sequences increases. In this paper, we propose a novel method to directly align the two heterogeneous sequences. We use the symmetric K–L divergence to measure the similarity of the elements in the two heterogeneous sequences. Based on the similarity matrix, the global sequence alignment is accomplished by a dynamic programming method called the Needleman–Wunsch algorithm [14], by which the computational complexity is dramatically reduced.

2. Face detection and track linking

The faces in the videos are detected by a cascade object detector in the Computer Vision System Toolbox of MATLAB. We divide the video into shots based on the differences of HSV features between neighboring frames. The faces which are continuous in position and scale within a shot are connected as a face track. The tracks whose average sizes of the bounding boxes are less than 55×55 pixels are regarded as false detection and deleted. Since faces in a track in neighboring frames are usually similar, we uniformly sample one face in every five consecutive frames in a track to reduce the data volume.

A facial landmark detector [15] is employed to detect five facial landmarks on each face. Based on the landmark points, we rectify the original detected faces to a canonical pose with a normalized distance of 30 pixels between two eyes. The facial landmarks and face rectification are illustrated in Fig. 2. On the rectified and cropped face image, we extract the gray level feature from a pixels patch centered by each landmark point (i.e. 40×30, 30×40,

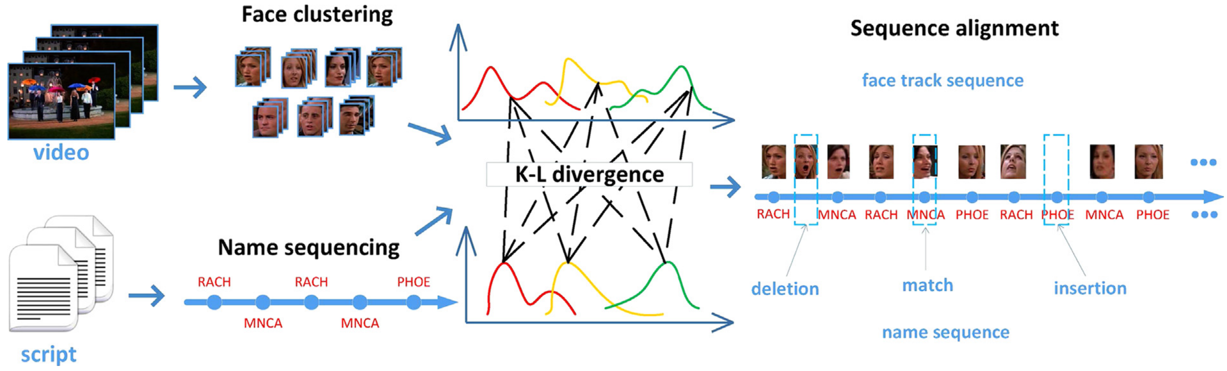

Fig. 1. Framework of face annotation based on face clustering and video/script alignment.

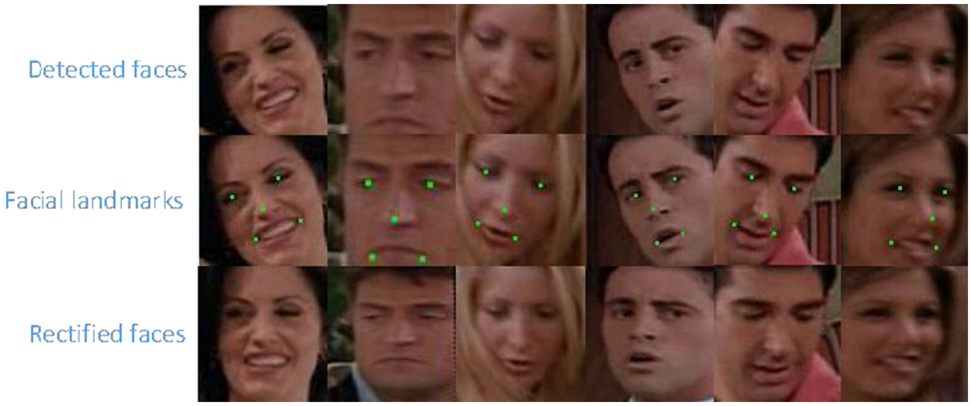

Fig. 2. Face rectification by the facial landmarks. The three rows from top to bottom are the original detected faces, the detected faces with landmarks and the rectified faces.