Neurocomputing 207 (2016) 240–249

Contents lists available at ScienceDirect

Neurocomputing

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

Face database generation based on text–video correlation

Dan Zeng [a], Yixin Bao [a,*], Ke Liu [a], Fan Zhao [a], Qi Tian [b]

[a] Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, Shanghai, China
[b] University of Texas at San Antonio, State of Texas, USA

ARTICLE INFO

Article history:
Received 5 December 2015
Received in revised form
15 March 2016
Accepted 2 May 2016
Communicated by Jinhui Tang
Available online 13 May 2016

Keywords:
Face database generation
DNN
Text–video correlation

ABSTRACT

The size of databases is the key to success to face recognition systems. However, building such a database is both time-consuming and labor intensive. In this paper, we address the problem by proposing a database generation framework based on text–video correlation. Specifically, visual content of a video can be presented as a character sequence by face detection, tracking and recognition, while text information extracted from subtitles and scripts provides complementary identity sequence. By correlating these two sequences, faces recognized can be refined without manual intervention. Experiments demonstrate that 90% of the human effort in face database construction can be reduced.

1. Introduction

Face recognition is a well-established field of research. A large number of algorithms have been proposed in the literature. Among them, deep neural networks [1–3] have yielded the best performance to date on LFW [4], YouTube Faces [5], and other challenging face benchmarks. This is partially due to the volume and complexity of the training data: DNN needs a huge amount of data to train millions of parameters and to generalize well to other datasets [6]. A relationship between training data size and face verification accuracy is shown in Table 1, which indicates how the performance improves with respect to the data size.

While building a large face database is in great need, it is often time-consuming and labor intensive. It is understandable since for a long time, datasets are collected in this way relying on extensive human labor. For instance, PubFig [7] spends a month collecting 6,500,000 user inputs from Amazon Mechanical Turk in order to label 60,000 images. Celebrity-1000 database [8] manually labels ~7000 videos of 1000 celebrities, and the entire process takes more than 1500 h.

While still necessary for the collection, we argue that faces in the images and videos can be labeled by some automatic mechanisms instead of by hand, e.g., face recognition algorithms. However, it is worth noting that face recognition alone is not so reliable for database generation, since recognition errors may occur. In this case, other supervise information is needed. Previous work has combined face detection with name annotation to identify people in images and videos [11–15], e.g., Everingham et al. [11] employ textual annotation for TV and movie footage, to automatically assign names to face images. Inspired by them, we utilize scripts and subtitles of videos to remove face recognition errors.

Specifically, subtitles record what is said and when, but not by whom, whereas scripts record who says what, but not when. Aligning these two textual annotations helps to extract who says what and when. Knowledge that a character is speaking suggests that the person may be visible in the video. On the other hand, face recognition helps to identify who is really visible in the video. Finally, correlation between the texts and the video can be established and made use of.

We formulate the problem as follows: given videos and their related scripts and subtitles, we aim at generating a large face database semi-automatically. By "semi-automatically", we mean keeping the amount of manual work to a minimum. Our solution includes three threads:

(1) Section 3 describes the subtitle–script alignment to obtain textual proposals of who says what and when. It helps to verify whether the recognized identity is correct or not.
(2) Section 4 describes video branch. A character sequence of who appears on-screen and when can be obtained by face tracking and identity recognition.
(3) Section 5 describes text and video correlation. We introduce timing projection method to accurately align two character sequences while retaining the time constraints within each other. This method helps to remove face recognition errors

* Corresponding author.
E-mail addresses: dzeng@shu.edu.cn (D. Zeng), elaine.bao@hotmail.com (Y. Bao), liuke@shu.edu.cn (K. Liu), shu_zfan@i.shu.edu.cn (F. Zhao), wywqtian@gmail.com (Q. Tian).

http://dx.doi.org/10.1016/j.neucom.2016.05.009
0925-2312/

and faces that pass through the validation can be stored in the database.

The whole framework is shown in Fig. 1. Experiments of the system are reported in Section 6, and further discussion is presented in Section 7. The contribution of this work lies in leveraging textual annotations of the videos to automatically correct face recognition outputs, and making face database generation more effective and efficient.

2. Related work

Face recognition: Face recognition has been long an active field of research [16,7,2,17,3], which can be further divided into two tasks: face identification and face verification. The main difference between them is whether we should know exactly the identity of the image. For face verification, we do not have to tell who is in the image, instead, the goal is to indicate whether two given images belong to the same person or not. This is quite different from our work, for database generation, we have to know the identities of the images and videos. On the other hand, face identification is more closer to our goal, but we put more emphasis on accuracy.

Pursuing higher accuracy is always a central topic in the field of face recognition. Deep convolutional neural networks have received considerable attention because of its state-of-the-art recognition accuracy [1–3]. These networks usually contain hundreds of millions of parameters. In training, millions of images are needed to prevent over-fitting.

Databases: A few databases are now available for training deep networks [4,5,18,19,20,21,30,31]. However, one common problem is that database construction involves huge amounts of labeling work. For example, image-based databases such as LFW, People In Photo Albums (PIPA) dataset [18] contain tens of thousands of images collected from thousands of people. Annotators are asked to filter plenty of albums and mark the regions of persons' heads in the photos. Video-based databases are built in similar ways. Youtube Faces DB [5] contains 3425 videos of 1595 different persons. Annotators are asked to filter unqualified videos, and label the identity of the person appearing in each video. CCV [19] database first collects 9317 consumer videos over 20 semantic categories from Youtube searches, and then ask Amazon Mechanical Turk users to watch the videos and classify them. The quality of consumer videos is usually low, which makes it more difficult for annotators.

Table 1
Face verification accuracy on LFW.

System	Accuracy (%)	No. of images/IDs
ConvNet-RBM [9]	92.52	87,628/5436
DeepFace [1]	97.35	4,400,000/4030
DeepID [10]	97.45	202,599/10,177
DeepID2+ [3]	99.47	290,000/12,000
FaceNet [2]	99.63	150,000,000/8,000,000

One main goal of our proposed work is to suggest a framework that can replace most of the current human effort in face dataset collection. A few recent approaches in this domain are closer to our current framework. Chen et al. [20] introduce a web video dataset for name–face association. Faces appearing in web videos are first associated greedily with names presented in the surrounding contexts, then annotators are asked to label each name–face pair as correct or not. In this work, the multiple classification problem of face identification is transformed into a binary classification problem, however, unlike ours, the name–face pairs generated by their method are still in need of manual labeling.

Li et al. [21] propose a method to automatically collect online pictures via Bayesian incremental model learning. First, a very small number of seed images are labeled by hand, and then a classifier is learned from these images to extract new images online. The newly collected images are added to the dataset, serving as new training data to improve the classifier, as a result, the database can be augmented iteratively. Similar to their work, our framework is also able to increase data volume automatically. Besides, we focus on video database generation, which is more diversified compared with image-based dataset, considering illuminating conditions, poses, motion blur, etc.

Text annotation: For more accurate face labeling, additional cues are being exploited. [12–14] aims at labeling "the faces in the news". The frequency of a person's visual appearance with respect to his name occurrence in transcripts is modeled as Gaussian distribution. Enlightened by them, we also use the timing pattern between names and faces to verify the identities of the faces.

Zhang et al. [15] propose a method to annotate faces in uncontrolled videos by global video/script alignment. A global sequence alignment algorithm is employed to find the most probable names for the faces according to their temporal distribution. The whole process is based on probability estimate, which is not reliable for database generation, as in our case.

In order to be more accurate, we make use of both scripts and subtitles of films and dramas, which has similarities with [11]. In that work, characters appearing on-screen are named automatically with speakers' names, which are generated by aligning subtitles and transcripts by dynamic time warping [22]. Much attention is focused on finding visible speakers in videos. Overall accuracy of naming characters is around 70%. In our case, we further embed a face recognition mechanism to improve the accuracy of visual information, and use subtitles and scripts as a verification signal to enhance face recognition results. By the end, our recognition accuracy based on text–video correlation exceeds 99%.

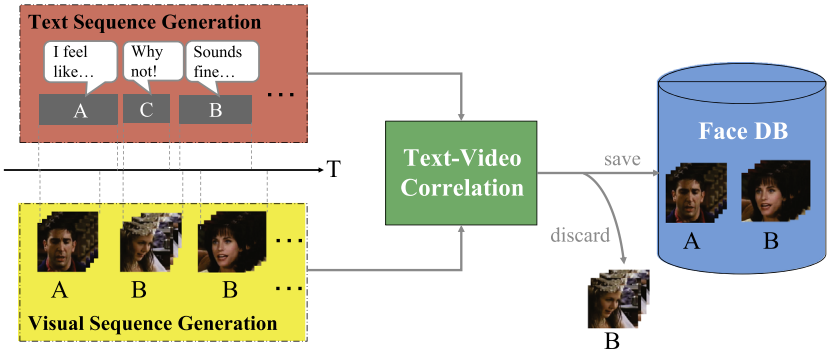

Fig. 1. Database generation framework. We correlate text and visual information to generate face database. Only faces with identities appearing both textually and visually can be confirmed, otherwise they will be discarded.