Face Recognition in Real-world Surveillance Videos with Deep Learning Method

Ya Wang, Tianlong Bao, Chunhui Ding, Ming Zhu
Department of Information and Technology
University of Science and Technology of China
Hefei, China
e-mail: wangyaya@mail.ustc.edu.cn

***Abstract*—Robust face recognition in real-world surveillance videos is a challenging but important issue due to the needs of practical applications such as security monitoring. While current face recognition systems perform well in relatively constrained scenes, they tend to suffer from variations in pose, illumination or facial expression in real-world surveillance videos. In this paper, we propose a method for face recognition in real-world surveillance videos by deep learning. First, a novel dataset from target real-world surveillance videos is constructed automatically and incrementally with the process of face detection, tracking, labeling and purifying. Then, a convolutional neural network with the labeled dataset is fine-tuned. On the testing dataset collected from the campus surveillance system, the network after fine-tuning achieves recognition accuracy of 92.1%, which obviously outperforms the network without fine-tuning, which returns a recognition accuracy of 83.6%.**

Keywords-face recognition; surveillance; deep learning

I. INTRODUCTION

Surveillance systems have been widely deployed in both public and private venues in recent years for the purpose of recognizing/verifying the subjects of interest in terms of security monitoring, and most of these systems are monitored by human operators. However, monitoring by humans inherently involves problems in reliability and scalability [1]. Hence, an intelligent approach to recognize the identities of different subjects by applying computer vision strategies has attracted the attention of researchers. Face recognition in surveillance videos is one of these strategies. Face recognition is natural, nonintrusive, and convenient to use, and face recognition systems have a wide range of applications in the public security, entertainment, attendance management, financial payment and other fields. While current face recognition systems perform well under relatively controlled environments [2], [3], they encounter great challenges when employed in practical surveillance videos due to the variations in image resolution, background clutter, illumination, facial pose and expression [4].

Deep learning is considered to be the greatest breakthrough in the field of artificial intelligence over the past decade, and has achieved great success in image classification [5], object detection [6], semantic segmentation [7], as well as face recognition [8]. Thanks to its deep architecture and powerful learning capacity, discriminative features can be learned through hierarchical nonlinear mappings, thereby significantly outperforming hand-crafted features. It is a very promising idea to apply deep learning methods to face recognition in real-world surveillance videos.

As we all know, to exploit deep learning methods to a new target domain, vast amounts of training data are critical. For example, Deep Face [8] collected 4.4M face images to do model training. However, labeling such a large scale dataset is a very tedious task. We cannot get a dataset of sufficient scale that is required for the depth of network as many ordinary research institutions or companies do. So, it is not realistic for us to re-train an entire model suitable for face recognition in real-world surveillance scenes from scratch with random initialization. Using the parameters of existing trained deep learnings model is another solution. While these models perform well on the public testing datasets, their performance is not satisfactory when applied to real scenarios. Taking VGG face model [9] as an example, when we tested the model on our real-world surveillance video dataset, the accuracy dropped to 83.6%, while its performance on the YTF dataset is 92.8%. To make the models more suitable to a real scenario, one of the solutions is fine-tuning the pre-trained models with the dataset collected from the target domain [10].

Motivated by this, we propose a method for face recognition in real-world surveillance videos by deep learning. The contributions of this paper mainly lie in two folds: First, an approach is proposed for creating a reasonable face dataset from target surveillance videos while requiring no or only a little help of human for annotation. Enough labeled data of the target domain is needed for fine-tuning the deep model. However, labeling such a dataset manually is tedious and costly. We construct a dataset novelly by a process of face detection, tracking, and graph-clustering. It is collected and labeled automatically. Then, a method for face recognition in real-world surveillance videos by fine-tuning the deep model with the new dataset is present.

In Section II, related work is discussed and Section IIIdescribes the details of the proposed method. In Section IV, the experimental results are presented. The paper is finally concluded in Section V.

II. RELATED WORK

In general, the methods of recognizing faces in videos can be divided into two categories: 1) Spatio-temporal recognition which is based on both static and dynamic information; 2) Frame-based recognition in which only one

978-1-5090-6238-6/17/$31.00 ©2017 IEEE

(or a few) static image(s) are available for recognition [11]. Videos can benefit face recognition by providing many frames that include abundant temporal dynamic information such as motion, pose and expression. Previous works show that it is necessary to make the most use of temporal information to improve the performance of face recognition. In [12], the information from different cameras and the temporal characteristics in a video sequence were incorporated by a dynamic Bayesian network. A hidden Markov model (HMM) was exploited to analyze the temporal clues over time in [13]. Some work has also been done to construct a 3D face model by using the dynamics in videos. For example, in [14], the author utilized a 3D head model which used features like hairline, to deal with large pose variations in head tracking and video-based face recognition.

When it comes to frame-based recognition, normally an automated face detector is first used to extract the faces from the video frames. Subsequently these faces are used for face recognition. It is similar to still-to-still image face recognition, which has been researched for such a long time that so many methods have been developed for it. According to whether the deep learning is used, the face recognition methods can be divided into traditional methods and deep learning based methods. Traditional methods rely on feature representation given by hand-crafted descriptors such as LBP [15], Gabor [16], eigenfaces [17] and so on. They perform well in constrained conditions, but encounter many challenges when used in unconstrained scenarios such as real-world surveillance videos due to the large variations in illumination, pose and low resolution. Many deep neural networks have been proposed for face recognition. Deep Face [18] employed an explicit 3D face model for alignment and represented the face with a nine-layer deep neural network, which was trained on a labeled dataset including 4 million facial images belonging to more than 4,000 identities. A supervised metric, i.e. theX2 similarity and the Siamese network, was used for the verification metric. Experiments showed the method reached an accuracy of 97.35% on the Labeled Faces in the Wild (LFW) dataset [18]. The Deep ID series of papers [19] [20] [21] increased the performance on LFW incrementally but steadily. In [8], the author trained a deep architecture, one that we call the VGG face model, using a triplet loss and achieved comparable state of the art results on the standard LFW and YTF [22] face benchmarks.

In this paper, the deep learning is employed to solve the problem of face recognition in real-world surveillance video. A relatively large scale face dataset is collected from our school monitoring system. Then, VGG face model is fine-tuned with the new dataset. Testing on our real-world surveillance videos dataset, the recognition accuracy of the fine-tuned model reaches 92.6%, improved by about 9% in comparison with the model without fine-tuning.

III. OUR METHOD

We propose a method for face recognition in real-world surveillance videos using deep learning. It consists of two parts. In the first part, a dataset was constructed by automatically collecting and labelling the data from real-world surveillance video. In the second part, we fine-tune the VGG face recognition model by our dataset. The two parts will be described in detail in the following subsections.

A. Dataset Collection and Labeling

As mentioned above, the dataset is important for fine-tuning. Since our target domain is real-world surveillance video, massive data from the target domain is necessary, including the number of identity classes and the amount of each identity class, to fine-tune our model. The dataset is established by four stages, as summarized in Table I.

- Stage 1: Rough dataset generation

Our goal is face recognition in real-world surveillance videos. One of the advantages of videos is that they involve many different person objects, so we can get enough identity classes from massive videos. Besides, each identity often appears in a series of frames. Thus enough face images for each identity are able to be obtained. These face images are various in illumination, pose, expression, and are even shaded to some extent. We build our dataset roughly utilizing the combination of face detection and face tracking. As showed in Fig. 1, when a video is input, the faces are detected by frames.

TABLE I. THE PROCEDURE OF DATASET COLLECTION AND LABELING

Stages	Aims	Methods
Stage1	Rough dataset generation	face detection and tracking
Stage2	Purification within each class	graph-clustering by VGG feature
Stage3	Purification between classes	similarity measure
Stage4	Filtering according to amount	quantity statistics

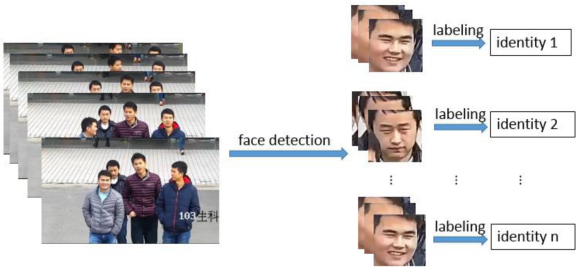

Figure 1. Rough dataset generation.

Figure 2. Some examples of an identity. The images with the red boxes are labeled incorrectly.

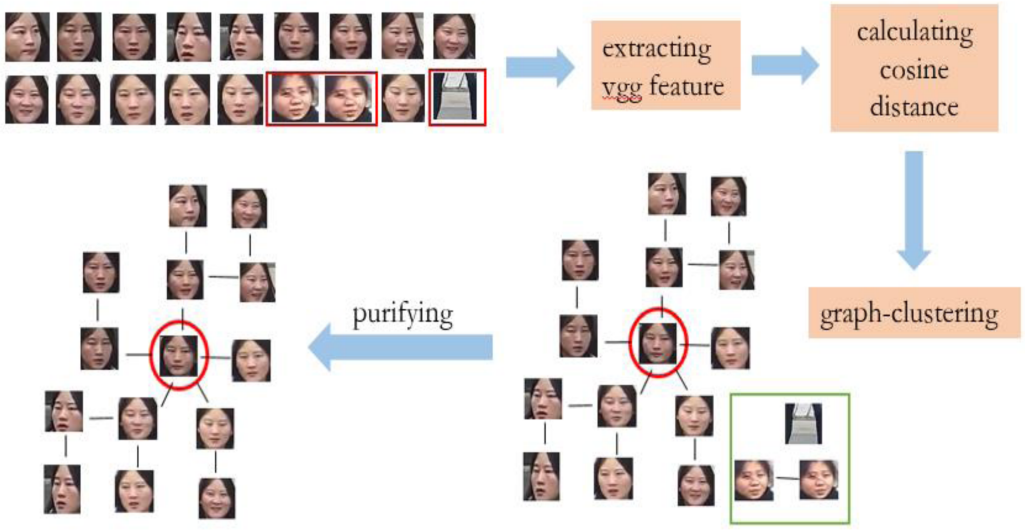

Figure 3. Some examples of an identity. The images with the red boxes are labeled incorrectly.

Once a face is detected, it will be tracked to get several face images of the identity and classed as an identity class.

- Stage 2: Purification within each class

The dataset created by stage 1 is mixed with some erroneous faces within each identity class set. Fig. 2 shows some examples of that. The aim of this stage is to remove them by the use of graph-clustering with the VGG face feature, which are extracted by the VGG face model. Specifically, the VGG feature is first extracted for each face image in the identity class set and the cosine distance between each face image is calculated. Then if the cosine distance between two images is under a tight threshold T, the two images are linked by an edge. By integrating all the edges and images, the graph is constructed. Considering the fact that most of the images within the identity class set are labeled correctly, which means they are related to a true identity, the image which owns the most edges is chosen as the center image and it represents the identity class. Naturally, the images linked to the center image directly or indirectly are selected to construct our new dataset. The procedure mentioned above is illustrated in Fig. 3.

- Stage 3: Purification between classes

There are some duplicate identities between identity class sets as showed in Fig. 4. They are cleaned by measuring the similarity of the center images. A threshold is set to decide whether the two identity classes are duplicated.

Figure 4. An example of duplicate identities. The first row is an identity and the second row is another identity. However, they are the same identity in fact. The images with red boxes are center images.

- Stage 4: Filtering according to amount

In this stage, identity classes for which there are not enough(less than 100) images are eliminated.

In summary, by a combination of face detection and tracking, purification within and between identity classes, and filtering according to amount, a dataset is established accurately which meets the requirement for fine-tuning.

B. *Fine-tuning*

In the fine-tuning stage, the VGG face model [8] is exploited as the pre-trained model. The VGG face model was trained on a large scale face dataset collected from the web (2.6M images, over 2.6K people). It comprises eight convolutional layers and three fully connected layer. Each of them are followed by one or more non-linearities such as ReLU and max-pooling [8]. The weights of the pre-trained VGG face model are fine-tuned by continuing the back-propagation with our dataset. Only the fully connected layers are fine-tuned with our new dataset to make the model more suitable to our target domain. This way of fine-tuning is accepted widely for the explanation that the features extracted by the earlier layers of a deep neutral network are generic and should be useful for most tasks, while the later layers are inclined to extract more specific details of the target domain.

IV. EXPERIMENTS

The effectiveness of our method is verified in two ways. More specifically, it is proved that an accurate dataset from real-world surveillance videos are able to be built by the method proposed with little participation from humans. Then, it is that the new dataset that the VGG face model can be fine-tuned and the performance on the target domain is improved.

The surveillance videos are collected for two consecutive weeks from our school surveillance system and our dataset is constructed following the method described in section 3. In the face detection stage, the haar face detector [23] is utilized

and KCF tracking method [24] is applied in the tracking step. Finally, the dataset includes 320 identity classes and over 80 face images for each identity class. Some examples are showed in Fig. 5. One of the evaluation indexes of the dataset is the purity and the purity of the dataset established by our method is 99.2%. The purity of the dataset is defined as follows:

$$\text{Purity} = N / M \quad (1)$$

N is the amount of images labelled correctly and M is the amount of all images.

Figure 5. Some face images collected from real-world surveillance videos.

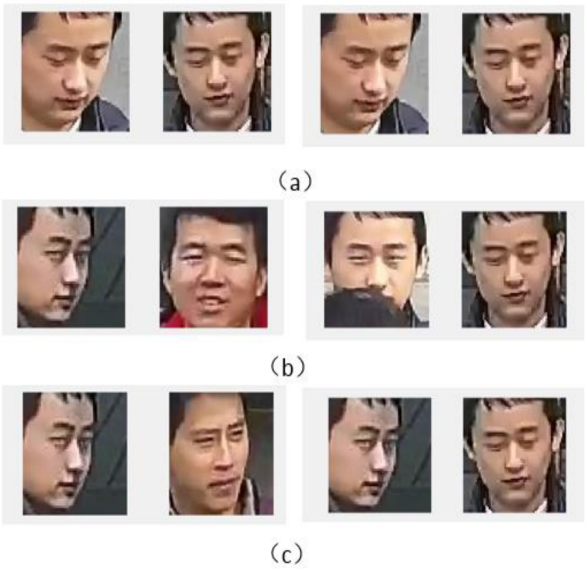

Figure 6. The left column is the result of the pre-trained model. The right column is the result of the fine-tuned model. The left images in each column are the testing images and the galleries are in the right of each column. As shown in (b)and (c), after fine-tuning, the recognition is obviously more accurate.

The training set (240 identity classes) and four testing sets (70 identity classes in each sets) are attained from the dataset. The performance of the pre-trained VGG face model is tested on the four testing sets and the average performance is chose as the final result. We use the model to extract the feature, and then the cosine distance and the nearest neighbor metric are used for recognition. In the part of fine-tuning, our mini-batch size is 64, and the learning rate for all convolutional layers and all full connected layers is 0.001 and 0.01 relatively. First the model is fine-tuned with 140 identity classes, and then enlarges the dataset to 240 identity classes, finally testing the model in the same way as the pre-trained model. The recognition accuracy results are shown in Table II and some image results are shown in Fig. 6.

TABLE II. THE RECOGNITION ACCURACY OF PRE-TRAINED AND FINE-TUNED MODELS

Modes	Identity class scale	Recognition rates
Pre-trained	——	83.6%
Fine-tuned1	140	91.4%
Fine-tuned2	240	92.1%

V. CONCLUSION

In this paper, a face recognition method for real-world surveillance videos was proposed by constructing a face dataset and fine-tuning the VGG face model. It is worth noting that the dataset was collected and labeled automatically from real-world surveillance videos with no or little participation of manual work for annotation, and, with our new dataset, it becomes possible to fine-tune the VGG face model to improve its performance in real-world surveillance scenarios. Also the VGG face model can be replaced by other probe models. On our testing dataset, the VGG face model after fine-tuning achieves a recognition rate of 92.1%, which obviously outperforms the VGG face model without fine-tuning, whose recognition rate is 83.6%.

REFERENCES

[1] D. Williams, "Effective CCTV and the challenge of constructing legitimate suspicion using remote visual images," Journal of Investigative Psychology and Offender Profiling, 2007.

[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," ACM computing surveys (CSUR), vol. 35, pp. 399-458, 2003.

[3] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J. Bone, "Face recognition vendor test 2002 results," Evaluation report, 2003.

[4] S. Patil and D. Deore, "Video-based face recognition: a survey," World Journal of Science and Technology, vol. 2, pp. 136-139, 2012.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91-99.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431-3440.

[8] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701-1708.

[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in British Machine Vision Conference, 2015, p. 6.

[10] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3208-3215.

[11] A. Hadid and M. Pietikainen, "From still image to video-based face recognition: an experimental analysis," in Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, 2004, pp. 813-818.

[12] L. An, M. Kafai, and B. Bhanu, "Dynamic bayesian network for unconstrained face recognition in surveillance camera networks," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 3, pp. 155-164, 2013.

[13] X. Liu and T. Cheng, "Video-based face recognition using adaptive hidden markov models," in Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, 2003, pp. I-340-I-345 vol. 1.

[14] M. Everingham and A. Zisserman, "Identifying individuals in video by combining'Generative'and discriminative head models," in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005, pp. 1103-1110.

[15] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in European conference on computer vision, 2004, pp. 469-481.

[16] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition," in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005, pp. 786-791.

[17] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of cognitive neuroscience, vol. 3, pp. 71-86, 1991.

[18] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst2007.

[19] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891-1898.

[20] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in Advances in Neural Information Processing Systems, 2014, pp. 1988-1996.

[21] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2892-2900.

[22] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 529-534.

[23] P. Viola and M. J. Jones, "Robust real-time face detection," International journal of computer vision, vol. 57, pp. 137-154, 2004.

[24] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, pp. 583-596, 2015.