Pattern Recognition 66 (2017) 153–159

Contents lists available at ScienceDirect

Pattern Recognition

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

A weakly supervised method for makeup-invariant face verification

Yao Sun[a], Lejian Ren[b], Zhen Wei[a], Bin Liu[c], Yanlong Zhai[b], Si Liu[a,*]

[a] State Key Laboratory of Information Security (SKLOIS), Institute of Information Engineering, Chinese Academy of Sciences, No. A89, Minzhuang Road, Haidian District, Beijing 100093, China
[b] School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China
[c] Moshanghua Tech Co. Ltd., Beijing, China

ARTICLE INFO

Keywords:
Face verification
Makeup-invariant
Weakly supervised method
Video context
Triplet loss function

ABSTRACT

Face verification, which aims to determine whether two face images belong to the same identity, is an important task in multimedia area. Face verification becomes more challenging when the person is wearing makeup. However, collecting sufficient makeup and non-makeup image pairs are tedious, which brings great challenges for deep learning methods of face verification. In this paper, we propose a new weakly supervised method for face verification. Our method takes advantages of the plentiful video resources available from the Internet. Our face verification model is pre-trained on the free videos and fine-tuned on small makeup and non-makeup datasets. To fully exploit the video contexts and the limited makeup and non-makeup datasets, many techniques are used to improve the performance. A novel loss function with a triplet term and two pairwise terms is defined, and multiple facial parts are combined by the proposed voting strategy to generate better verification results. Experiments on a benchmark dataset (Guo et al., 2014) [1] and a newly collected face dataset show the priority of the proposed method.

1. Introduction

Human face verification [2,3] has been extensively studied and has various practical applications. In human perception and psychology studies [4], it has been revealed that heavy makeup can significantly decrease the human ability of recognizing faces. As shown in Fig. 1, significant appearance changes can be observed for individuals with and without makeup.

Convolutional Neural Network (CNN) extracts features from the raw image data, and has achieved great success in many tasks such as image recognition [5–8], fashion [9–11] and general face recognition [12]. Although methods based on deep learning have achieved competitive results, there is a great difficulty to use CNN for our task, which needs to do face verification to those with heavy makeups. The difficulty is that it is hard to find or collect enough makeup and non-makeup face pairs for training the network.

In this paper, due to the lack of large amount makeup and non-makeup datasets, we propose a weakly supervised makeup-invariant method for face verification. Our method takes advantages of large amount of free video contexts from the Internet, and only needs small sets of labelled makeup and non-makeup images to fine-tune. The pipeline of our method is demonstrated in Fig. 1. First, a large amount of free videos are downloaded from the web. Next, face detection and tracking are used to generate the *positive pairs*. By saying positive pairs, we mean that the faces belong to the same persons, while the persons appearing in the successive frames are usually regarded as identical. Negative pairs can be selected by some strategies or simple random selected. Then we can get triplet pairs from the video contexts. These triplets are used to pre-train a model. With the pre-trained weights, we only need to collect small amount of before–after makeup faces in the next step. By fine-tuning the model on these small amounts of makeup and non-makeup images, we can obtain a face verification model robust to makeups. To avoid overfiting in the fine-tuning stage, we use many techniques in our network, including using larger features, extra pairwise losses, mirror images, and local facial parts. These techniques will be detailed in Section 3. Experiments on a benchmark dataset [1] and a newly collected face dataset show the advantage of the proposed method.

Our major contributions can be concluded as follows.

(1) We propose a weakly supervised method for face verification that is robust to cosmetic changes and achieves state-of-the-art performance.
(2) We propose a deep framework for makeup-invariant face verifica-

* Corresponding author.
E-mail addresses: sunyao@iie.ac.cn (Y. Sun), renlejian@outlook.com (L. Ren), zhen.wei@hotmail.com (Z. Wei), liubin@dress-plus.com (B. Liu), ylzhai@bit.edu.cn (Y. Zhai), liusi@iie.ac.cn (S. Liu).

http://dx.doi.org/10.1016/j.patcog.2017.01.011
Received 15 July 2016; Received in revised form 6 January 2017; Accepted 7 January 2017
Available online 10 January 2017

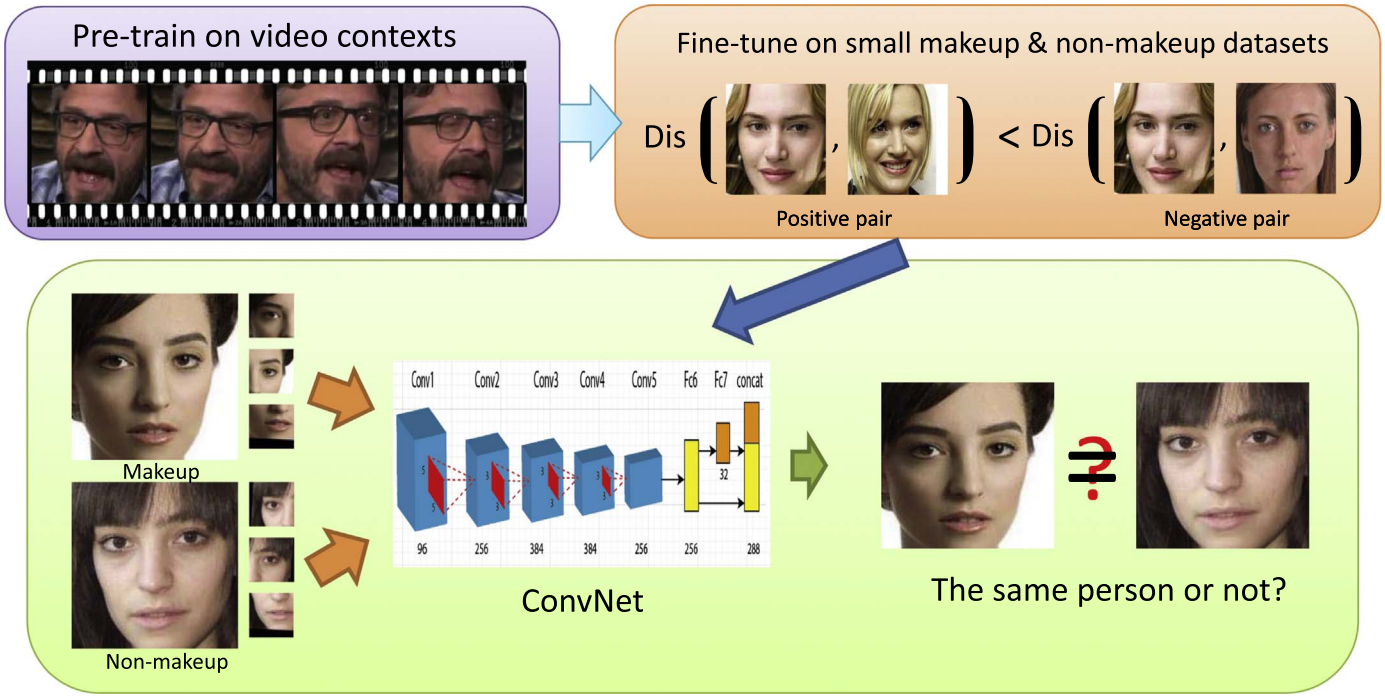

Fig. 1. A weakly supervised method for face verification.

tion, which has two distinguishing properties, i.e. (a) it utilizes freely available videos for pre-training and (b) multiple facial parts are combined to achieve better performance.

(3) We collect a large scale video face dataset and a before–after makeup pair dataset, which can be used as the benchmark in the further studies.

2. Related works

Face verification: Recently, there are increasing interest and great progress in the face verification task. Taigman et al. [13] proposed DeepFace which carefully designed both the alignment step and the representation step by employing explicit 3D face modelling in order to apply a piecewise affine transformation, and derived a face representation from a nine-layer deep neural network. Sun et al. [3] proposed a hybrid convolutional network (ConvNet)-Restricted Boltzmann Machine (RBM) model for face verification in wild conditions. Sun et al. [14] proposed DeepID which can be effectively learned through challenging multi-class face identification tasks, whilst they can be generalized to other tasks (such as verification). Sun et al. [15] proposed a Deep IDentification-verification features (DeepID2). The face identification task increases the inter-personal variations by drawing DeepID2 features extracted from different identities apart, while the face verification task reduces the intra-personal variations by pulling DeepID2 features extracted from the same identity together. In [2], Schroff et al. proposed a Facenet system for face recognition and clustering. A triplet network was used to train the network and several sample selecting strategies were discussed as well. There are some other works that address face verification tasks under specific conditions, such as with occlusions [16,17] or in videos [18]. However, these works are not specifically designed for makeup invariant face verification task.

Weakly supervised learning for face recognition: Face data in images and videos are of great volume on the Internet. They can be easily obtained through video websites and social websites. Previous work [14] has proved that getting more face data and more identities involved helps to improve recognition performance.

However, manually labelling such great number of face data is laborious and impractical. Recent researches begin to address more importance on weakly supervised labels. Rim et. al. [19] leveraged weak labelled data for face recognition based on probabilistic graphical models. Chen and Deng [20] built up a challenging unlabelled database and proposed an efficient Self-Learning DCNN structure (SL-DCNN) to handle weakly supervised training for face recognition. In this paper, we make use of unlabelled video context data to generate weakly supervised labels for model pre-training.

Makeup studies: Recently, there are more works focusing on the makeup related studies, such as makeup transfer [21–23] and makeup recommendation [24].

A dual attributes approach [25] was proposed to learn facial attributes in makeup and non-makeup faces separately, and face matching uses the semantic-level attributes to reduce the influence of makeup on low-level features. Another approach [26] is to preprocess face images with a self-quotient image technique to reduce makeup effects before face matching. However, these methods cannot reduce the makeup influence significantly. Guo et al. [1] proposed performing correlation mapping between makeup and non-makeup faces on features extracted from local patches.

3. Methods

A triplet network is presented in our methods and is illustrated in Fig. 2. Based on this network, we take three stages to obtain the final face verification results. Firstly, we pre-train the network on video contexts which are easy to get from public available videos. Next, we fine-tune the pre-trained models on makeup and non-makeup images, and we also fine-tune on several parts of these images in the second stage. In the last stage, by using a *voting* approach, we summarize the verification results obtained both from the whole face images and the parts of the faces, and give a final decision on whether the input two images belong to the same identity.

Details of the method are given in the following.

3.1. The triplet network

Our goal is to learn a discriminative feature representation so that

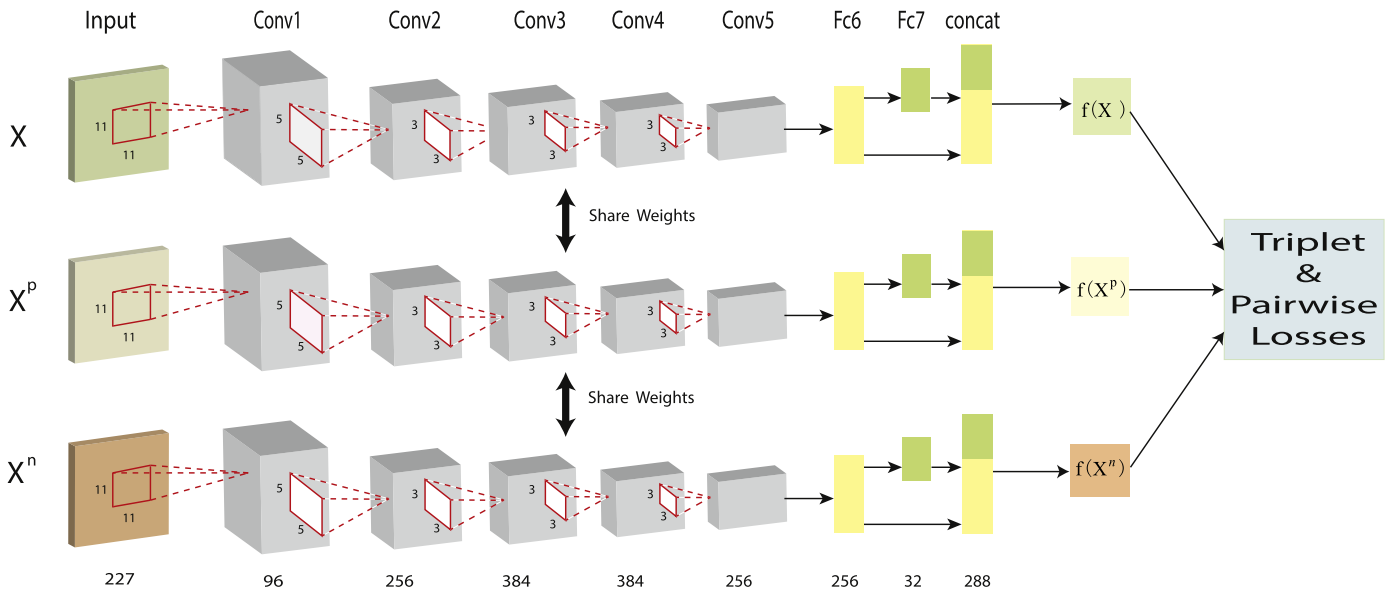

Fig. 2. The framework of the proposed triplet convolutional neural network. There are three input images. They go through three convolutional neural networks respectively but share the same parameters. At the final layer, the features $f(X)$, $f(X^p)$, $f(X^n)$ are extracted, and triplet and pairwise loss functions are used.

the similarity of two faces of the same person is higher than the faces from different people. To this end, we design a triplet convolutional neural network. The triplet network consists of three face inputs, where the first and second faces form a positive pair, and the first and third faces form a negative one. The construction of positive and negative pairs is further elaborated in Section 3.2. These faces are input into three parallel networks. All these three networks share the same parameters.

The detailed network parameters are shown in Fig. 2. The backbone network is based on the AlexNet [5] framework. We use the same convolutional layer structure settings of the AlexNet. Two fully connected layers follow the convolutional layers. Since the number of makeup and non-makeup images used for training is quite limited, to avoid the network from overfitting, the neuron numbers of these two fully connected layers (Fc6 and Fc7) are truncated to 256 and 32 respectively. However, the 32-dimension feature vectors in the last fully connected layer are not sufficient to represent all the characteristics of the input images, so we concatenate the features of the Fc6 and Fc7 layers and obtain a 288-dimension feature vector in the final feature space. Then loss functions are defined in this 288-dimension feature space. In our experiments, using the concatenated features (Fc6 + Fc7) instead of the features in the last layer (Fc7) does improve the performance of our method. We think this is because the 288-dimension feature vector space is more representative.

3.2. The loss functions and verifications

The loss functions: Three loss functions are used in our network. One is the triplet rank loss and the other two are pairwise losses. All these loss functions are based on the standard cosine distances of two vectors. For two n-dimension vectors $A = (a_i)$ and $B = (b_i)$, the cosine distance of them is

$$d(A, B) = 1 - \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \times \sqrt{\sum_{i=1}^{n} b_i^2}}.$$

The first loss function is the *triplet rank loss function*. Let (X, X^p, X^n) be a triplet of the input network, where the faces in the images X and X^p belong to the same person and the face in X^n is different. For convenience, we say X^p is a *positive sample* of X and X^n is a *negative sample*. We denote $f(X)$, $f(X^p)$ and $f(X^n)$ as the 288-dimension feature vectors computed by our network for the input images X, X^p and X^n respectively. Parameters of the networks are omitted for simplification. We hope that the distance between $f(X)$ and $f(X^p)$ is smaller than the distance of $f(X)$ and $f(X^n)$, i.e.

$$d(f(X), f(X^p)) + \alpha < d(f(X), f(X^n)), \tag{1}$$

because the faces in X and X^p belong to the same person, where α is a margin that is a constraint on positive and negative pairs. For this purpose, we define the triplet rank loss as:

$$l_{rank}(f(X), f(X^p), f(X^n)) = \max\{0, d(f(X), f(X^p)) - d(f(X), f(X^n)) + \alpha\}. \tag{2}$$

In our experiments, we set α=0.1.

The triplet rank loss function encourages the image $f(X)$ to become more similar to $f(X^p)$ than $f(X^n)$, which has been extensively exploited, e.g. [2]. But we note that the triplet rank loss function does not constrain how similar $f(X)$ is to $f(X^p)$ and how different $f(X^n)$ is from $f(X)$. To constrain the distances $d(f(X), f(X^p))$ and $d(f(X), f(X^n))$ more tightly, we use two pairwise loss functions. That is, we define

$$l_{pos}(f(X), f(X^p), f(X^n)) = \max\{0, d(f(X), f(X^p)) - (\delta - \alpha/2)\}, \tag{3}$$

and

$$l_{neg}(f(X), f(X^p), f(X^n)) = \max\{0, (\delta + \alpha/2) - d(f(X), f(X^n))\}, \tag{4}$$

where δ is a threshold given in advance. We use $\delta = 0.5$ in the training stage of our experiments. From the comparisons in Table 2, we find that using these two pairwise loss functions does improve the performance, and this is probably because using more loss functions prevents the model from overfitting.

The final loss is

$$l = l_{rank} + \lambda_{pos} \cdot l_{pos} + \lambda_{neg} \cdot l_{neg}. \tag{5}$$

Here, λ_{pos} and λ_{neg} are balance weights. We use $\lambda_{pos} = \lambda_{neg} = 1$ in our experiments. Remark that this loss is summed up on all triplet pairs in a mini-batch. Stochastic Gradient Descent (SGD) is used to minimize this loss.

Triplet pairs selection: In the triplet ConvNet training, triplet pairs are the input. A triplet pair includes an image X, its positive sample X^p, and its negative sample X^n. During pre-training stage, following the work [27], a positive pair (X, X^p) is generated by faces from the same video. Negative pairs (X, X^n) are faces from randomly selected different videos. For fine-tuning stage, a positive pair is given by dataset

Fig. 3. Parts of the images.

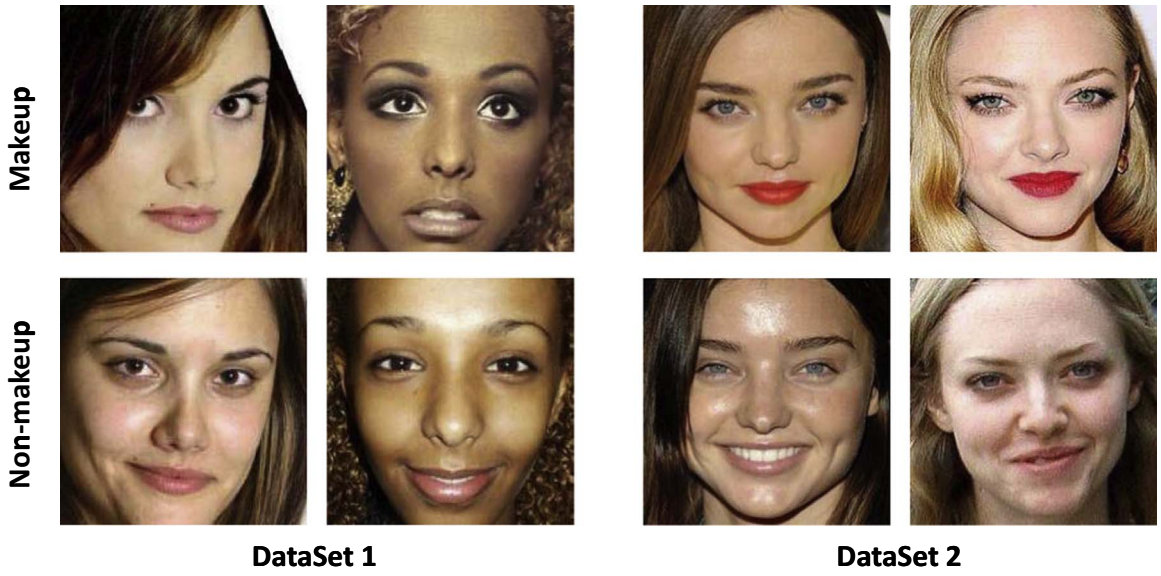

Fig. 4. Makeup and non-makeup images in datasets.

annotations. Negative pairs are makeup and non-makeup faces from different persons. Here we use *hard negative selection* strategy [27] to replace *random selection* strategy after training for 10 epochs. Positive–negative ratios in both stages are 1:4.

Face verification: To verify whether a makeup face X and a non-makeup face Y belong to the same person, we extract the features $f(X)$ and $f(Y)$ first. If the distance $d(f(X), f(Y))$ is smaller than a given threshold, we then think the persons in X and Y are identical; otherwise, X and Y are regarded as different persons. The threshold is decided in the verification stage.

To improve the performance, besides training the ConvNet on the whole face images, we also train this triplet ConvNet on parts of the images. Some facial parts of the images are shown in Fig. 3.

To integrate the results from the ConvNets trained on parts, we use a voting method. For a makeup face X and a non-makeup face Y, we compute the features $f(X)$ and $f(Y)$ from the whole faces, and we also extract the features $f_i(X_i)$ and $f_i(Y_i)$ from parts, where X_i and Y_i are parts of X and Y. To decide whether X and Y belong to the same person, we define

$$\sigma(X, Y, \gamma) = \begin{cases} 1, & \text{if } d(f(X), f(Y)) \le \gamma, \\ -1, & \text{if } d(f(X), f(Y)) > \gamma, \end{cases}$$

and calculate

$$v = \lambda \cdot \sigma(X, Y, \gamma) + \sum_i \sigma(X_i, Y_i, \gamma_i).$$

We believe X and Y are the same person if $v > 0$; otherwise, X and Y are different persons. In experiments, we set $\lambda = 1.1$ to emphasize the affects of the whole faces, which also ensures $v \neq 0$.

We use three parts of the images in our experiments. These three parts contain the left eye, the right eye, and the mouth respectively. We believe our method can be further improved by using more parts.

4. Experiments

4.1. Experiment Setting

Dataset1 for makeups and non-makeups: To qualitatively evaluate

our method, we use the dataset collected in [1]. There are 1002 face images of 501 female adults. Each individual has two images, one for makeup and the other for non-makeup. To the best of our knowledge, this dataset is the largest one we can access.

Dataset2 for makeups and non-makeups: We also collect some makeup and non-makeup images from the Internet. We get 203 females and their 406 makeup and non-makeup images.

Some images in the above two datasets are shown in Fig. 4.

Dataset for pre-training: Since the makeup and non-makeup images are quiet limited, directly training on these sets leads to overfitting easily. So we pre-train our ConvNet on other related datasets. Following the work [27], we take advantage of the public available video sources on the Internet. To construct the pre-trained dataset, we downloaded videos containing faces from http://YouTube.com using face related keywords, such as talk show and smile. Around 30,000 video clips are obtained.

Then we use a detection-tracking strategy to crop out faces. Video clips are first partitioned with scene boundaries using algorithm in [28]. Then the detector pre-trained on AFLW dataset [29] is used to perform face detection on these clips. Only detection results with confidence over 0.95 are used. When the detector cannot find a face, we use an off-the-shelf tracker [30] to perform tracking and only retain the tracked faces with high confidence.

By using triplet pairs selection strategy (Section 3.2), we generated 11,250 positive pairs of faces. Some of the faces are shown in Fig. 5.

Baselines: We consider three baseline methods. There is one method using the traditional non-deep learning method and two methods using unsupervised trained models as initial weights.

Baseline1: In paper [1], the authors developed a face recognition system that is robust to facial makeup. Correlation mapping between makeup and non-makeup faces on features extracted from local patches is performed. A complete face verification system is developed based on their makeup detection results.

Baseline2: The network proposed in [27] is a triplet network for unsupervised learning from videos. We fine-tune their proposed network on the "color model" (from *github*). On performing face verification, we first find out the best threshold, and then regard two faces as the same person if the cosine distance of the extracted feature vectors are smaller than the threshold.

Baseline3: We fine-tune the AlexNet by using the model provided in [31]. The authors presented another good method for self-supervised training. The model for AlexNet-like networks is downloaded from the project cite of this paper. When fine-tuning the AlexNet, we replace the Fc8 layer by a triplet loss layer to do face verification. The verification method is the same as Baseline2.

Evaluation metrics: We follow the evaluation method of [1]. A five-fold cross validation is used. In the testing stage of face verification, there are 100 and 40 pairs of positive faces in each round of the two datasets respectively. We randomly generate 100 and 40 negative pairs from the 100 and 40 positive pairs. The accuracy is calculated on all these positive and negative pairs.

Implementation details: All our experiments are on Caffe [32]. The parameters of the triplet networks are shown in Fig. 2. The initial learning rate is 0.001, and decreases by a factor 10 after every 50 epochs. The size of mini-batch is 100. Random triplet selection strategy is used for the first 10 epochs both in the pre-training and fine-tuning stages, and the hard triplet selection strategy is used afterwards. We use Gaussian initialized parameters for the pre-training stage and stop after 200 epochs. We fine-tune another 200 epochs on the before–after makeup datasets using the pre-trained weights.

Using mirror images: To improve the robustness of the network, we also use the mirrors of input images more aggressively. Let X be an input image, and $\overline{X}$ be its mirror (in horizontal). We input both the triplet (X, X^p, X^n) and its mirror $(\overline{X}, \overline{X^p}, \overline{X^n})$ to the network at the same time, and compute the loss $l([f(X); f(\overline{X})], [f(X^p); f(\overline{X^p})], [f(X^n); f(\overline{X^n})])$ instead of $l(f(X), f(X^p), f(X^n))$, where $[A; B]$ means the concatenation of A and B. That implies, when using mirror images, the final features space is 576-dimension and will provide better representations of the original images.

4.2. Experimental results and analysis

We compare our weakly supervised face verification method with three baseline methods in Table 1. The top three rows are baseline methods, and the bottom three rows are our method using different techniques. The row "Our method w/o video context" shows the performance of our method training directly on the makeup and non-makeup datasets with Gaussian initialized parameters. "Our method+video context" refers to our method trained only on the whole face images by using pre-trained weights. Our complete method "Our method+video context+parts" also fine-tunes on parts of the faces and uses the voting approach for the final verification.

Fig. 5. Positive pairs collected from video clips. The top row shows the faces of individuals in selected frames, and the bottom row shows their corresponding faces after 10 frames in the same clips.

Table 1
Comparisons with baseline methods.

Method	Acc. on Dataset1 (proposed in [1]) (%)	Acc. on Dataset2 (newly collected)
Baseline1 [1]	80.5	–
Baseline2 (fine-tuned on [27])	78.5	62.1%
Baseline3 (fine-tuned on [31])	73.8	64.2%
Our method w/o video context	73.3	Do not converge
Our method+video context	81.1	65.1%
Our method+video context+parts	**82.4**	**68.0**%

Table 2
Effects of techniques.

Method	Acc. on Dataset1 (proposed in [1]) (%)	Acc. on Dataset2 (newly collected) (%)
Only video contexts	68.6	55.5
Fc7+l_{rank}	77.9	61.0
Fc6+Fc7+l_{rank}	79.0	63.1
Fc6+Fc7+(l_{rank}, l_{pos}, l_{neg})	79.8	64.3
Fc6+Fc7+(l_{rank}, l_{pos}, l_{neg})+mirror	**81.1**	**65.1**
Our method+video contexts [part1]	74.2	64.2
Our method+video contexts [part2]	76.6	62.1
Our method+video contexts [part3]	74.0	63.0
Our method+video contexts + parts	**82.4**	**68.0**

From the comparisons, we can see that our weakly supervised method outperforms the non-deep learning method (Baseline1) and the two weakly supervised methods (Baseline2 and 3) in both datasets. For the three baseline methods, Baseline1 performs better than Baseline2 and 3. We think this is because the limited sizes of the makeup and non-makeup datasets make the deep learning systems fail to find out more representative features. All methods perform better on Dataset1 than Dataset2, because Dataset2 contains even fewer images. Our method also suffers from limited size of datasets. With Gaussian initialized parameters, our method perform poorer than all baselines. However, when using pre-trained weights from video contexts, our method outperforms all others and achieve accuracy 81.1% and 65.1% on the two datasets. Since our network has similar structures to the two deep learning baselines, we believe it is the techniques we used that improve the performance. Using facial parts even advances the proposed method.

In Table 2, we compare the effects of techniques used in the proposed method. "Only video contexts" means we use the pre-trained model trained on video contexts for face verification directly. In method "Fc7+l_{rank}", features are only extracted from the Fc7 layer and only the classical triplet rank loss function is used. We concatenate features from Fc6 and Fc7 layers in "Fc6+Fc7+l_{rank}". We use two more loss functions in "Fc6+Fc7+(l_{rank}, l_{pos}, l_{neg})". The features of mirror images are used in "Fc6+Fc7+(l_{rank}, l_{pos}, l_{neg})+mirror". The following three rows report the accuracies of our method fine-tuned on parts of the images. The last row is the performance of our complete method.

From Table 2, we can see that all the techniques can improve our method. The pre-trained models on video contexts are not suitable for the face verification with heavy makeups, because the makeups are really very confusing. After fine-tuning on the makeup and non-makeup datasets, the performances are significantly improved. Using 288-dimension features (Fc6+Fc7) instead of 32-dimension feature (Fc7) make the feature vectors more representative and hence improves the performance. More loss functions decrease the risk of overfitting of the model, and using mirror images even extends the

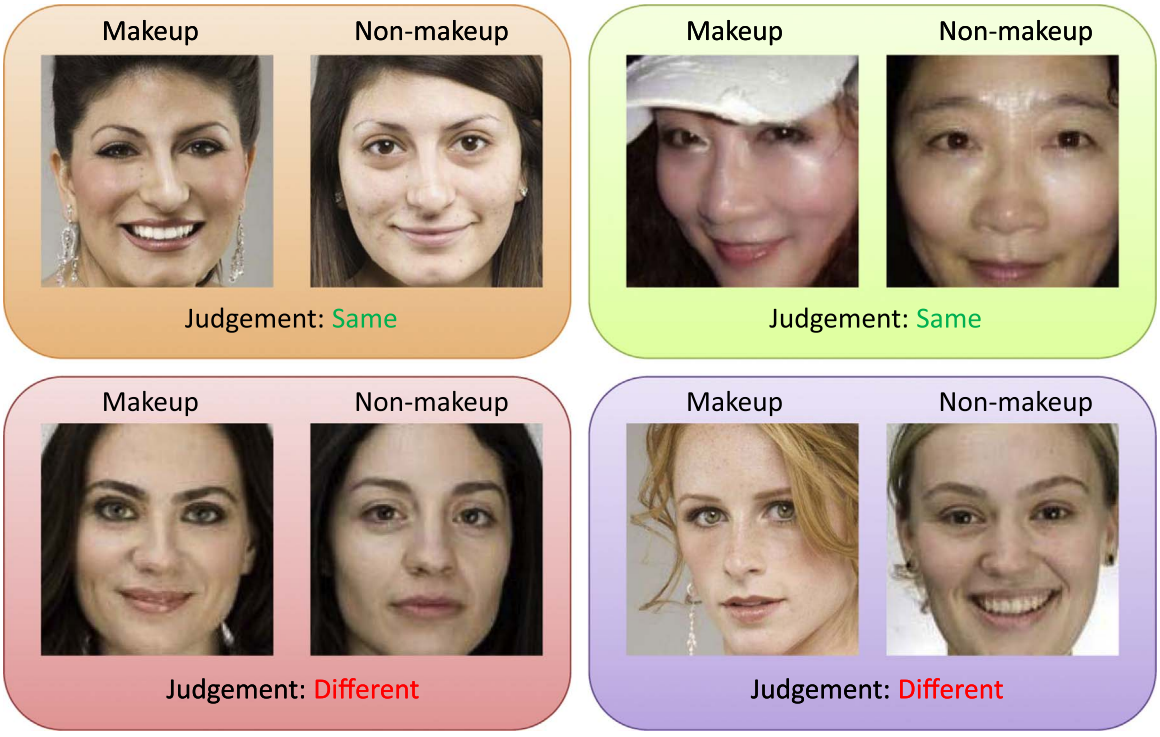

Fig. 6. Some qualitative results.

feature spaces larger. Both of these techniques make the proposed method even better. Note that, since facial parts only contain partial information of the whole face, our method performs a bit poor on each single part. But the combination of the whole face and all parts results can improve the final verification accuracy.

The method "Fc7+l_{rank}" is similar to the one proposed in [2], but does not obtain the best performance. We think there are two possible causes. First, the network developed in [2] aims to do general face recognition but not specialize to face verification with heavy makeups. Second, since the number of makeup and non-makeup images used for fine-tuning are quite limited, more techniques should be used to prevent overfitting.

Some verification results are also presented in Fig. 6 for qualitative evaluation.

5. Conclusion

In this paper, we propose a new weakly supervised method for face verification, which is robust to cosmetic changes. Video contexts are used for unsupervised pre-train and improve our method significantly. We also present many techniques for further improvements, including using concatenation of features, pairwise loss functions, mirror images, and voting approach on models trained on parts. We also collect a large scale video face dataset and a before–after makeup pair dataset for further studies. We believe that our method can be further improved by using more video contexts information and more parts of the faces.

Acknowledgments

This work was supported by National Natural Science Foundation of China [Nos. 61602037, 11301523, and 61572493, Grant U1536203].

References

[1] G. Guo, L. Wen, S. Yan, Face authentication with makeup changes, IEEE Trans. Circuits Syst. Video Technol. 24 (5) (2014) 814–825.
[2] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: CVPR, 2015, pp. 815–823.
[3] Y. Sun, X. Wang, X. Tang, Hybrid deep learning for face verification, in: ICCV, 2013, pp. 1489–1496.
[4] G. Rhodes, A. Sumich, G. Byatt, Are average facial configurations attractive only because of their symmetry?, Psychol. Sci. 10 (1) (1999) 52–58.
[5] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS, 2012, pp. 1097–1105.
[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, arXiv preprint arXiv:1409.4842.
[7] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, arXiv preprint arXiv:1502.01852.
[8] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, S. Yan, Matching-CNN meets KNN: quasi-parametric human parsing, in: CVPR, 2015, pp. 1419–1427.
[9] S. Liu, Z. Song, G. Liu, C. Xu, Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set, in: CVPR, 2012, pp. 3330–3337.
[10] S. Liu, T.V. Nguyen, J. Feng, M. Wang, S. Yan, Hi, Magic closet, tell me what to wear!, in: ACM MM, 2012, pp. 619–628.
[11] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, S. Yan, Fashion parsing with weak color-category labels, IEEE Trans. Multimedia 16 (1) (2014) 253–265.
[12] X. Wu, R. He, Z. Sun, A lightened CNN for deep face representation, arXiv preprint arXiv:1511.02683.
[13] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in: CVPR, 2014, pp. 1701–1708.
[14] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10, 000 classes, in: CVPR, 2014, pp. 1891–1898.
[15] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: NIPS, 2014, pp. 1988–1996.
[16] S. Zhang, R. He, Z. Sun, T. Tan, Multi-task convnet for blind face inpainting with application to face verification, in: ICB, 2016, pp. 1–8.
[17] J. Qian, L. Luo, J. Yang, F. Zhang, Z. Lin, Robust nuclear norm regularized regression for face recognition with occlusion, Pattern Recognit. 48 (10) (2015) 3145–3159.
[18] Z. Huang, R. Wang, S. Shan, X. Chen, Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning, Pattern Recognit. 48 (10) (2015) 3113–3124.
[19] D. Rim, M.K. Hasan, F. Puech, C.J. Pal, Learning from weakly labeled faces and video in the wild, Pattern Recognit. 48 (3) (2015) 759–771.
[20] B. Chen, W. Deng, Weakly-supervised deep self-learning for face recognition, in: ICME, 2016, pp. 1–6.
[21] D. Guo, T. Sim, Digital face makeup by example, in: CVPR, 2009, pp. 73–79.
[22] K. Scherbaum, T. Ritschel, M. Hullin, T. Thormählen, V. Blanz, H.-P. Seidel, Computer-suggested facial makeup, Comput. Graph. Forum 30 (2) (2011) 485–492.
[23] W.-S. Tong, C.-K. Tang, M.S. Brown, Y.-Q. Xu, Example-based cosmetic transfer, in: CGA, 2007, pp. 211–218.
[24] L. Liu, J. Xing, S. Liu, H. Xu, X. Zhou, S. Yan, Wow! You are so beautiful today!, ACM Trans. Multimed. Commun. Comput. Appl. 11 (1s) (2014) 20.
[25] L. Wen, G. Guo, Dual attributes for face verification robust to facial cosmetics, Comput. Vis. Image Process. 3 (1) (2013) 63–73.
[26] C. Chen, A. Dantcheva, A. Ross, Automatic facial makeup detection with application in face recognition, in: ICB, 2013, pp. 1–8.
[27] X. Wang, A. Gupta, Unsupervised learning of visual representations using videos, in: ICCV, 2015, pp. 2794–2802.
[28] E. Apostolidis, V. Mezaris, Fast shot segmentation combining global and local visual descriptors, in: ICASSP, 2014, pp. 6583–6587.
[29] K. Martin, W. Paul, P.M. Roth, B. Horst, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, in: ICCVW, 2011, pp. 2144–2151.
[30] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2012) 1409–1422.
[31] C. Doersch, A. Gupta, A. Efros, Unsupervised visual representation learning by context prediction, in: ICCV, 2015.
[32] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093.

Yao Sun is an Associate Professor in the Institute of Information Engineering, Chinese Academy of Sciences. He received his Ph.D. degree from the Academy of the Mathematics and Systems Science, Chinese Academy of Sciences.

Lejian Ren is an Undergraduate in the Internet of Things Engineering from the School of Computer Science and Technology, Beijing Institute of Technology.

Zhen Wei received the B.S. degree in Computer Science and Technology from Yingcai Honors School, the University of Electronic Science and Technology of China, Chengdu, China. He is now a Graduate Student at the Institute of Information Engineering, the Chinese Academy of Science.

Bin Liu received his Bachelor of Electronic Engineering from Southwest University of Science and Technology and Master of Science degree in Electrical Engineering from Shanghai Jiao Tong University, China in 2008 and 2011 respectively. His recent research interests include use of deep learning methodology for modelling, classification and tracking.

Yanlong Zhai is an Assistant Professor in the School of Computer Science, Beijing Institute of Technology. He was a Visiting Scholar in the Department of Electrical Engineering and Computer Science, University of California, Irvine. He received Ph.D. degree from Beijing Institute of Technology. His research interests include distributed and parallel computing.

Si Liu is an Associate Professor in Institute of Information Engineering, Chinese Academy of Sciences (CAS). She was a Research Fellow in Learning and Vision Research Group at National University of Singapore. She obtained Ph.D. degree from Institute of Automation, CAS. Her research interests include object categorization, object detection, image parsing and human pose estimation.