Contents lists available at ScienceDirect

Pattern Recognition

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

Fine-grained face verification: FGLFW database, baselines, and human-DCMN partnership

Weihong Deng*, Jiani Hu, Nanhai Zhang, Binghui Chen, Jun Guo

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Keywords:
Fine-grained visual recognition
Face verification
Labeled face in the wild
Deep learning

ABSTRACT

As performance on some aspects of the Labeled Faces in the Wild (LFW) benchmark approaches 100% accuracy, there is an intense debate on whether unconstrained face verification problem has already been solved. In this paper, we study a new face verification problem that assumes the imposter would deliberately seek a people with similarly-looking face to invade the biometric system. To simulate this deliberate imposture attack, we first construct a Fine-Grained LFW (FGLFW) database, which deliberately selects 3000 similarly-looking face pairs within original image folders by human crowdsourcing to replace the negative pairs of LFW. Our controlled human survey reports 99.85% accuracy on LFW, but only 92.03% accuracy on FGLFW. As the algorithm baselines, we evaluate several state-of-the-art metric learning, face descriptors, and deep learning methods on the new FGLFW database, and their accuracy drops about 10–20% compared to the corresponding LFW performance. To address this challenge, we develop a Deep Convolutional Maxout Network (DCMN) which aim to tolerate the multi-modal intra-personal variations and distinguish fine-grained localized inter-personal facial details. The experimental results suggest that the proposed DCMN method significantly outperforms current techniques such as Deepface, DeepID2, and VGG-Face. Fusion of the scores of our proposed DCMN to that of human operators notably boost the verification accuracy from 92–96%, suggesting that human-algorithm partnerships are promising to detect the similarly-looking deliberate impostors.

1. Introduction

As the explosion of ubiquitous biometric data, there has been a significant progress in improving face recognition accuracy due to big data driven machine learning methods. After the eras of subspace learning [1,2] and sparse representation [3,4], deep learning technique shows extraordinary effectiveness to solve the unconstrained face recognition problem. Many deep learning methods [5,6] have reported nearly saturated accuracy on the standard Labeled Faces in the Wild (LFW) benchmark. LFW provides a “same/different” benchmark which addresses the face recognition problem as a non-class-specific similarity problem and which is different from more traditional multi-class classification problem [7]. The rationale is that such a benchmark requires that methods learn to evaluate the similarity of faces rather than be able to recognize particular faces. The power of the same/not-same formulation is in diffusing a multi-class task into a manageable binary class problem. Moreover, by removing all the test subject from the training, LFW encourage learning face similarity, rather than the distinguishing features of particular face. Thus, the benchmark aims to gain a generalization ability which is not limited to a predefined set of classes. The reported accuracy is a concise index on face recognition performance regardless of the number of candidate classes.

While the performance of LFW benchmark approaches 100% accuracy, the community is still addressing only part of the overall face verification problem. After inspecting the LFW databases, one can identify a main limiting factor for its unconstrained face verification task: *almost all the negative face pairs are quite easy to distinguish*. The negative pairs are randomly selected from different individual, and it is common that two random individuals have large differences in appearance. Many face pairs even have different genders. Thus, verification is, by its nature a problem in which many examples are very easy with large inter-class variance, because the collection of LFW database is based on the assumption of *random imposter attack*. For practical usage, however, it is likely that a desperate impostor may attempt to spoof a genuine user by seeking a similarly-looking people. This real-world difficulty is important to the ubiquitous biometric killer applications such as the face pay in Internet finance, and video surveillance based person re-identification. Unfortunately, this common and realistic challenge for face biometric has not been explicitly evaluated or addressed before.

* Corresponding author.
E-mail address: whdeng@bupt.edu.cn (W. Deng).

http://dx.doi.org/10.1016/j.patcog.2016.11.023
Received 16 July 2016; Received in revised form 27 November 2016; Accepted 28 November 2016

To fill up this blank, we reinvent the LFW database to explicitly evaluate the face verification accuracy under *desperate imposture attack*. The new database, called Fine-Grained LFW (FGLFW), is collected by crowdsourcing efforts that seek 3000 similarly-looking face pairs (300 pairs per fold) to replace the random negative pairs of LFW. The positive pairs of FGLFW are identical to those of LFW database. To distinguish FGLFW from LFW, the prefix "Fine-Grained" suggests that the difference between inter-class face pairs is so tiny that even the *human operators* would feel compelled to make fine-grained inspect on the localized facial features during the verification. There are three motivations behind the construction of FGLFW benchmark as follows.

1. Continuing the intensive research on LFW with more realistic consideration on deliberate imposture, and fostering research on the fine-grained face recognition in unconstrained images. The challenge of LFW benchmark focuses mainly on suppressing the large intra-class variance, such as poses and lighting, while FGLFW benchmark emphasizes both the large intra-class variance and the *tiny inter-class variance* simultaneously.
2. Constructing a moderately "difficult" database for evaluating the level of security provided by human operators, and stimulating research on the verification cases that are difficult for human to recognize, with an ambitious goal that design algorithms to help human to make reliable verification judgement.
3. Maintaining the protocols, dataset size, and the image ensemble of LFW database to encourage fair and meaningful comparisons, and allowing easy comparison and replication of results. The image ensemble of the FGLFW database is identical to LFW, so that one can study the fine-grained face verification performance based on the mature model for LFW.

To the best of our knowledge, FGLFW is the first benchmark that a large number of human perceptually similar face pairs are intentionally integrated into the evaluation of the face recognition system. Our controlled human survey yields 99.85% accuracy on LFW but only 92.03% accuracy on FGLFW, which suggests that random imposters are too "easy" but deliberate imposters are moderately "difficult" for human operators to detect. As the algorithm baselines, we evaluate several metric learning, advanced face descriptors, and deep learning methods on the new FGLFW database, and their accuracy drops about 10–20% compared to the LFW performance.

To boost the fine-grained face verification performance, we develop a Deep Convolutional Maxout Network (DCMN) which aims to tolerates the multi-modal intra-personal structures and, at the same time, discriminates fine-grained localized inter-personal facial features. Although the proposed DCMN achieves 91% accuracy, which is better than the other architectures such as Deepface, DeepID2 and VGG-Face, it still cannot by itself address fine-grained face verification problem. From the experimental results we can conclude that the fine-grained face verification is clearly beyond the current state-of-the-art and further research is required to address this more realistic and challenging setting. Finally, we attempted to fuse the similarity score from human and DCMN. Surprisingly, human-algorithm fusion *cuts the error rate by nearly a half* in comparison to the human verification. This complementary ability clearly suggests that it is promising to study the algorithms that enhance human's ability to recognize the deliberate imposter.

This journal paper is an extended version of the conference paper [8] of ICB 2016. In the paper, the new contents include the detailed discussion about fine-grained face verification problem, the comparative human survey results on LFW and FGLFW, a new DCMN method for the fine-grained face verification and the comparative study on deep learning methods, and the score fusion of human operators and DCMN to boost the performance.

2. Background

In 2005, Ferencz et al. [9,10] developed a method for deciding whether two images represented the same object. They presented this work on data sets of cars and faces, and hence were also addressing the face verification problem. To make the problem challenging for faces, they used a set of news photos collected as part of the Berkeley Faces in the Wild project [11,12] started by Tamara Berg and David Forsyth. These were news photos taken from typical news articles, representing people in a wide variety of settings, poses, expressions, and lighting. These photos proved to be very popular for research, but they were not suited to be a face recognition benchmark due to the more than 10% noisy labels and large numbers of duplicates. Eventually, after manual data cleaning and new protocols designing, the refined data were released as "Labeled Faces in the Wild" in 2007 [7].

Since that time, hundreds of papers have been published that improve upon this benchmark in some respect. A remarkably wide variety of innovative methods have been developed to overcome the challenges presented in this database, and Learned-Miller et al. presented a comprehensive survey on the great progress [13]. Under "unrestricted with labeled outside data" protocol where unlimited external data could be used to train the classifier, several approaches have reported over 99% accuracy. In particular, the highest reported accuracy on LFW described by a peer-reviewed publication stands at 99.63%, by Schroff et al. [5], reporting only 22 errors on the entire test set of 6000 image pairs. Under this protocol, as LFW test is too easy, the testing process may become an exercise in "tuning" existing algorithms, which makes distinguishing between algorithms nearly impossible.

Recently, several large-scale database, such as IJB-A [14], FaceCrub [15], CASIA-WebFace [16], and Megaface [17], have been designed to study large-scale face verification and identification problem. They advocate to measure the performance using performance criteria that are more strict than that of LFW. For verification, the verification rate at 0.1% false acceptance rate. For identification, rank-1 recognition accuracy on a gallery of thousands or millions of people is reported. Intuitively, these experiments would also involve many comparisons between the test image and similarly-looking gallery faces. However, such "similarity" is defined by algorithms, rather than human operators. In addition, while reporting more realistic performance, these new databases lose the virtues of LFW as the easy-to-use, low barriers to entry. In contrast, we aim to explicitly evaluate the algorithm to distinguish the similarly-looking negative face pairs that are "difficult" for human to verify. The related fine-grained recognition techniques is demanding in the biometric applications. At the same time, we design the database by strictly following the protocols of LFW so that researchers need not to do any change when transferring to the new benchmark. These two characteristics of make the proposed FGLFW database totally different from the recently proposed benchmarks.

Similar to the fine-grained face recognition, identical twin recognition problem [18] also require algorithms to distinguish the tiny inter-class variations. However, there are at least two differences between fine-grained face verification and identical twin recognition. Firstly, only 0.4% people have identical twin, but our study finds that more than half people can find out at least one similarly-looking people among a population of hundreds. Deliberate attack is much more common in the biometric system, and thus "fine-grained" face recognition problem is a very common challenge that has not been systematically studied before. Secondly, multi-modality biometrics are often required to distinguish identical twin [19,20], because there are possibly not enough discriminative information solely by the facial appearance at all. In contrast, FGLFW database encourage the algorithms to accurately distinguish the deliberate imposters by only the facial appearance. From the aspect of data source, FGLFW is totally different from the previous twin databases. To the best of our knowledge, all identical twin databases are collected at twins festivals [18–

21], where both the number of twin pairs and the image sessions are limited. Images of these databases may not diverse enough to model the real-world intra-class variations. In contrast, FGLFW database contains the photographs collected from the web designed for studying the problem of unconstrained face recognition.

Fine-grained face verification is a similar, but not the same, task to the fine-grained visual categorization [22] in the computer vision community. The fine-grained visual categorization problem asks us to distinguish subordinate-level categories such as husky dog and poodle dog from each other. Typical approaches to fine-grained visual categorization are based on detecting and extracting features from particular parts of the objects [23,24]. *Face recognition is an extreme case of fine-grained visual categorization in which the "subcategories" are individual instances* [25]. Similar to the fine-grained approaches, the best face recognition methods, such as high-dimensional LBP, extract features from locations determined by finding facial landmarks such as the corners of the eyes [26]. In this sense, many face recognition algorithms are the analogy to fine-grained categorization methods. In contrast, compared with the common face verification that address mainly large intra-class variations, such as pose, illumination, and expression, the proposed *fine-grained face verification task emphasizes the tiny inter-class difference in the deliberately selected similarly-looking face pairs*. As we shown in the experiment section, commonly used face recognition (fine-grained classification) strategy is not sufficient to solve the problem. Recent progresses on fine-grained visual categorization [27–30] may inspire future research ideas on fine-grained face verification.

3. From LFW to fine-grained LFW

In a biometric scenario, a basic assumption of LFW, as well as the other face biometric benchmarks, is that.

- *The imposter would randomly choose a people to invade the biometric system.*

This assumption requires the benchmarks to evaluate a massive number of negative face pairs ensure the security of the system. In practice, however, although the imposters would be much fewer than the genuine users, they are not likely to intrude into the system in a random manner. To address this issue, the key assumption of FGLFW is that.

- *The imposter would deliberately seek a people with similarly-looking face to invade the biometric system.*

We called this assumption the *deliberate imposture* assumption. To the best our knowledge, no previous research has studied this important issue on either human operators or machine algorithms.

3.1. Deliberately seeking imposters by crowdsourcing

Besides the deliberate imposture, we would like design FG-LFW to guide the new algorithms to focus on the issues that current methods could not addressed. Therefore, we design a crowdsourcing task to select the negative face pairs that are "difficult" to distinguish for both human labelers and recognition algorithms. Crowdsourcing process is effective to imitate the imposter who aims to select a similarly-looking face to intrude into the biometric system. We also limit the selection within each image fold of LFW database with a population of hundreds. We adopt the triplet labeling to avoid the inter-personal difference of the perception of similarity degree, and label the similarity in a relative (comparative) way.

To select the potential similar face pairs, we apply cosine similarity measure of the well-established DeepID2 descriptor [6] that is highly discriminative for face verification. For a triplet (Q, A, B), we regard Q as the query image. Supposing the similarity of Q and A is s, we choose a face of which similarity to Q is around s as B. And s is controlled to be higher than a specified threshold, i.e. 0.85, to avoid most triples being annotated as both A and B being dissimilar if s is too small.

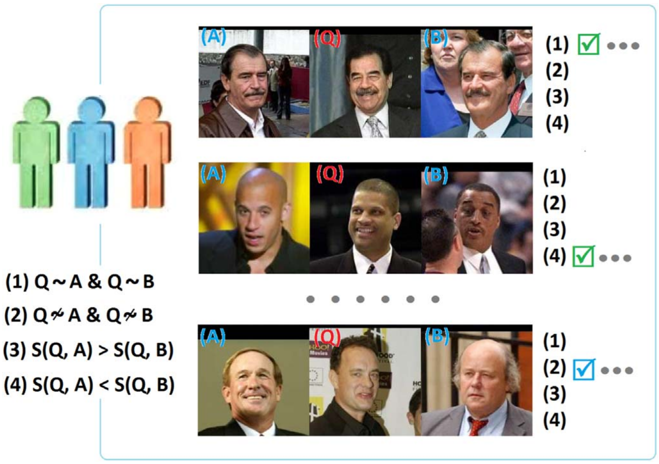


Fig. 1. An illustration of the crowdsourcing labeling of the triplet (A, Q, B) that selects the human perceptually similar face pairs to simulate deliberate imposter. Triplet labeling ranks the similarity in a relative way, and thus avoids the inter-personal difference of the perception of similarity degree.

LFW contains 13,233 face images of 5749 persons collected from Internet with large variations, in which View 2 defines 10 disjoint subsets of image pairs which are suitable for cross validation. Each subset contains 300 positive pairs and 300 negative pairs. In fact, the 10 subsets are organized by their identity, i.e. each identity only appears once in certain subset. Based on it, LFW database has been divided into 10 separate image folds. The triplet (Q, A, B) are chosen from each fold separately, and we get a total number of about 200,000 triplets. Specifically, in the online crowdsourcing as illustrated in Fig. 1, a face triplet is randomly displayed on the screen and the participants are asked to describe the triplet using one of the four choices as follows:

(1) The query Q is similar to both A and B;
(2) The query Q is not similar to both A and B;
(3) The query Q is more similar to A than B;
(4) The query Q is more similar to B than A.

Undergraduate students from the School of Information and Communication Engineering at the Beijing University of Posts and Telecommunication volunteered to participate in these experiments in exchange for a research credit in a pattern recognition course. 127 voluntary students were invited to attend this triplet labeling work, and over 600,000 labels of triplet have been collected in a period of one month.

For the selection of pairs, we split every triplet into two pairs, i.e. (Q, A) and (Q, B). The image pair receive a "similar" annotation if it is labeled as "more similar than the other pair" or "similar to the query" during its corresponding triplet annotations. Since each triplet are annotated for about three times and each pair may appear in different triplets in our annotation program, we respectively calculate the frequencies of each pair being annotated as being similar and dissimilar. With these frequencies, we can compute each pairs probability of being similar. Probability of pairs being annotated for few times often contains much noise. We use Laplace smoothing to alleviate it, and the human *perceptual similarity* measure of a image pair is defined as

$$p_i = \frac{c_i^+ + \delta}{(c_i^+ + \delta) + (c_i^- + \delta)} \tag{1}$$

where (c_i^+ denotes the number of "similar" annotation, (c_i^- denotes the

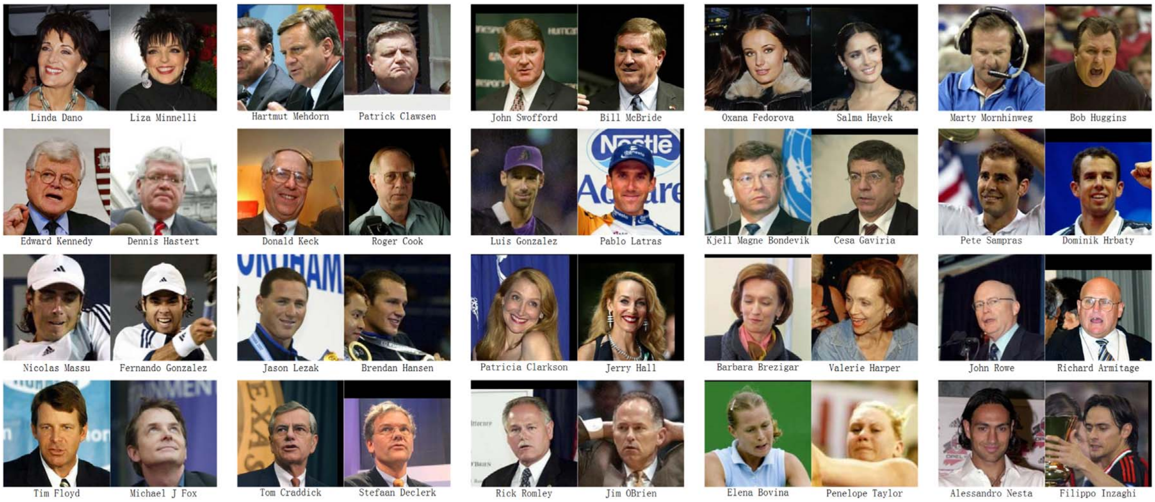


Fig. 2. Example negative face pairs selected by crowdsourcing triplet labeling. 3000 similarly-looking negative face pairs are selected to replace the negative pair of LFW, generating the new FGLFW database. Note that these challenging face pairs are manually selected within each image-fold in LFW consisting of only 400–450 people. Thus, one should aware that the deliberate imposture by similar face pairs would frequently happen in real-world face biometric system.

number of “dissimilar” annotation, $\delta = 2$ is the smoothing parameter. In each image folder, face pairs are sorted by corresponding perceptual similarity and first non-repetitive 300 pairs are selected as the deliberate imposter face pairs. Eventually we keep the 3000 positive pairs of original LFW protocol and replace the negative pairs with our selected 10×300 visually similar pairs. Example negative face pairs in the new database are shown in Fig. 2, which are confusing even for human operators even after careful inspection. Note that there are only hundreds of people in each fold of the database, which means the such similarly-looking face pairs can be seek in a small population in practice.

It should be noted that the deliberate imposture attack defined in the FGLFW database are much more common than the well-known “identical twin” attack [18], because there are only 0.4% people have identical twin, but our study find that more than half people can find out at least one people with similarly-looking faces among a population of hundreds, as illustrated in Fig. 3. Due to unexpected frequency of similarly-looking faces, it is possible that a large number of imposters have successfully cheated the human operators or automatic biometric system. By manual selection of these face pairs, we construct FGLFW database to facilitate further study on the distinguish of these similar, but not the same, faces. This may be important to close the large gap between the reported performance on benchmarks and performance on real world tasks.

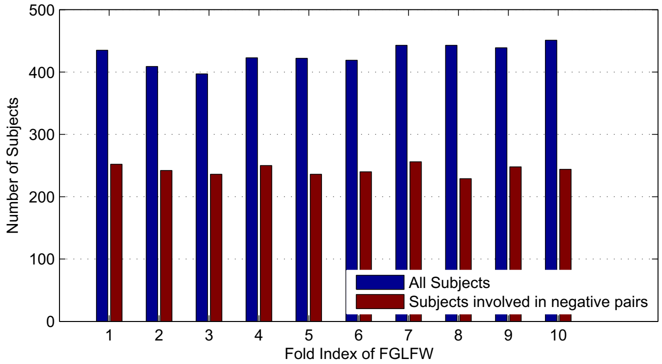


Fig. 3. In each fold of FGLFW database, the number of subjects involved in the negative pairs are more than half of the population. This means that more than half people can seek a similarly-looking faces in the population of hundreds.

3.2. Full compatibility with LFW

Following the data partition of LFW database, Fine-Grained LFW database defines 10 disjoint subsets of image pairs which are suitable for cross validation, where the image ensemble of each subset is identical to the LFW database. Each subset contains 300 matched face pairs and 300 negative face pairs. The 300 matched pairs of each fold are identical to those of LFW, and the 300 negative pairs are selected by crowdsourcing from the same image ensemble of the corresponding fold of LFW. In this manner, the FGLFW database is fully compatible to LFW for all the six protocols and evaluation criteria. Originally, there were two distinct protocols described for LFW, the image-restricted and the unrestricted protocols. The unrestricted protocol allows the creation of additional training pairs by combining other pairs in certain ways. (For details, see the original LFW technical report [7].) As many researchers started using additional training data outside LFW to improve performance, new protocols were developed to maintain fair comparisons among methods. These protocols were described in an updated technical report [31]. The current six protocols are:

1. Unsupervised.
2. Image-restricted with no outside data.
3. Unrestricted with no outside data.
4. Image-restricted with label-free outside data.
5. Unrestricted with label-free outside data.
6. Unrestricted with labeled outside data.

Since FGLFW only modifies the negative face pairs defined in standard protocol, the original training and testing paradigms of LFW can be directly used.

4. Human survey

To validate our database, we have conducted a human survey on the FGLFW challenge. The survey results were used for the following purposes: (1) Test the difficulty posed by the FGLFW challenge to human operators. (2) Provide a convenient means of comparing human performance to that of the existing state of the art. (3) Verify whether it is possible to design algorithms to help human operators to reject the deliberate imposters.

4.1. Controlled human survey on FGLFW

Different from the crowdsoucing survey using Amazon Mechanical

Turk [32], to ensure the reliability of the human labels, 27 registered students of our laboratory are invited to participate our *controlled* human survey, who were informed that their verification accuracy would be recorded and reported. In addition, after every ten face pairs are completed, the verification accuracy is updated and displayed in the screen in order to continuously remind the participants to focus attention on the task. In this manner, the result is a good simulation on the security level provided by responsible human operators. Besides, the participants of our survey are the Chinese students of 20–25 years old, who were probably not familiar with the appearance of most of the identities, since most LFW faces are outdated foreign politicians, sports figures, and actors in the last decade. That is, our survey on human performance roughly control that participants had no prior exposure to the people pictured in the test sets, conforming to the LFW protocol.

The human survey was conducted on all the 6000 face pairs of FGLFW benchmark, which are randomly partitioned to 3 non-overlapped subset of 2000 face pairs. Each participant viewed 2000 randomly selected pairs and was asked to rate his or her confidence that each of these pairs represents the same face on a 1–7 likelihood scale. Participants had unlimited time to enter a response for each pair, with images remaining on the screen until a response was entered. We have collected 54,000 answers from the 27 participants on the 6000 pairs, and each pair has been rated by 9 users.

Fig. 4 shows the ROC performance of all the 27 participants, which shows that most participants obtain about 80–90% accuracy on FGLFW face-matching task. Clearly, the human verification accuracy is far from perfect, and deliberate imposters are difficult to detect if a visually-similar face fairs are intentionally selected. In real-world security tasks, there may have being a large number deliberately seeking imposters have successfully cheated the human operators. Designing algorithms to improve the human security strongly motivates further research into fine-grained face verification methods. To measure the human performance, participant votes for each pair are treated as independent experts and their mean likelihood answer is the finally human score. Finally, the human performance reach to 92.03% by fusing the scores of all the 27 participants, suggesting that about 8% face pairs from deliberately seeking imposters cannot detected by human operators.

On average, the participants took as long as about 6.8 s to judge for one face pair. In contrast, for the previous face-matching experiment, studies had reported that human generally takes less than 2 s to compare a face pair [33]. Our participants feedback that they have to take a long time to make the verification judgement by carefully comparing the fine-grained details of the face pairs, such as the specific shape of eyebrow, nose, ear, or hairline. This recognition process is totally different from the daily face recognize by a glance at the holistic appearance. *Modeling and imitating human's fine-grained inspection on similar-looking face may be an interesting issue for further research* [22].

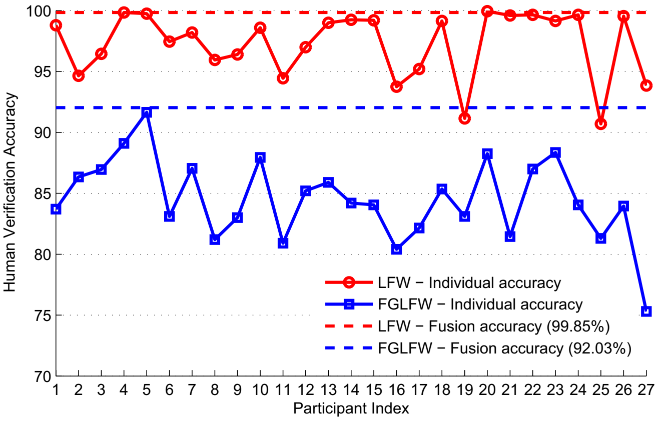


Fig. 5. Comparative verification accuracy of the 27 participants in our controlled survey on LFW and FGLFW databases. The fusion accuracy yielded by the average individual score is also illustrated. While the accuracy of LFW is nearly perfect, there are about 8% face pairs of FGLFW database cannot be discriminated by human.

4.2. Contrast survey on LFW

The controlled human survey reports 92.03% verification accuracy on the FGLFW database, which is much lower than that the 99.20% accuracy reported on LFW database [32]. In one hand, the large gap between databases suggests that the verification task defined in FGLFW is much more challenging than the LFW database even for responsible human operators. In the other hand, it is also possible that our survey suffer from the own-race effect that the faces of one's own race are better recognized than faces of other less familiar race. To eliminate this possibility, we run a *contrast survey* on the 6000 face pairs of LFW by the same 27 participants and identical rules. Astonishingly, as shown in Fig. 5, 11 out of the 27 participants achieve over 99% accuracy and the sum fusion of all the labeled likelihood yields a near perfect 99.85% accuracy, as shown by the black line in Fig. 6. The results clearly confirm that our survey do reflect the

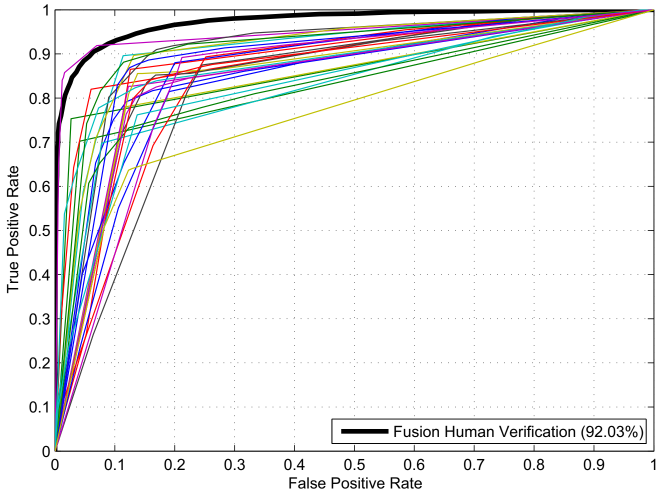


Fig. 4. The ROC curves of the 27 participants of our human survey on FGLFW. Each user viewed 2000 randomly selected pairs and was asked to rate his or her confidence that each of these pairs represents the same face on a 1–7 likelihood scale. The (thick black) fusion ROC curves is generated by the mean likelihood answer.

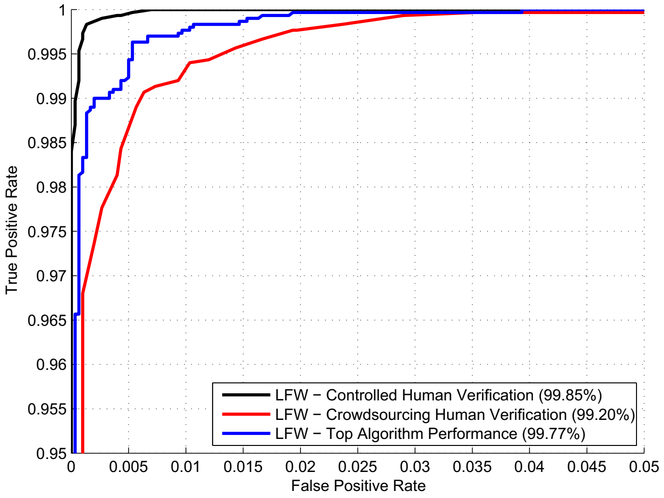


Fig. 6. The ROC curves on LFW database by human surveys and top algorithm. Controlled human survey is detailed in this paper. Crowdsourcing human survey was conducted by Kumar et al. [32]. Top algorithm performance is downloaded from http://vis-www.cs.umass.edu/lfw/results.html.

reasonable ability of human face recognition.

We acknowledge that our survey done by Chinese students is bound to suffer for own-race effect. However, our controlled survey indeed achieved higher accuracy than that reported by Kumar et al. [32]. A possible explanation on our perfect accuracy is that our restricted settings, such as the controlled participants and the interface with accuracy alert, have avoided most careless mistakes happened in the unrestricted crowdsroucing. The participants feedback that, since most negative pairs of LFW look apparently different, they could make very quick judgements on LFW database. That means there is no difficulty for human verification, even considering the own-race effect. Interestingly, the correlation coefficient of individual accuracy between LFW and FGLFW reaches 0.59. As illustrated in Fig. 5, the good recognizer in FGLFW tends to be a good recognizer in LFW.

Therefore, the accuracy of our survey is more close to the upper bound of the human face recognition ability. The 99.85% human accuracy clearly suggests that the negative face pairs of LFW may be too “easy” to test the human face recognition, on which a large proportion of errors in previous human survey [32] might be caused by careless mistakes of the participants. Removing the biases caused by careless mistakes, human is still better at that face verification task than leading algorithms such as FaceNet [5]. The generalization ability of the algorithms may be over-estimated in several recent studies that claimed to surpass human performance. As illustrated in Fig. 6, both human and algorithm can yield near perfect accuracy on LFW database, and thus further exploration would become the fine tune of a small number of unrepresentative face pairs, rather than solving the problem itself. Instead, FGLFW provides a moderately “difficult” benchmark for evaluating the level of security provided by both algorithms and human operators.

5. Deep convolutional max-out network

This section introduces a new DCMN model, which incorporates the very deep architecture with small convolution filters, the maxout units, and the verification supervised signal to address the fine-grained face verification task.

5.1. Motivations and components

In FGLFW database, the technical challenge of fine-grained verification comes from both the tiny inter-class difference and the large intra-class variance. In this situation, discriminative feature might only be learned in a multi-layer highly-nonlinear manner. Therefore, we attempt to incorporate some recently proposed components, such as very small convolution filters, the maxout units, and the verification supervised signal for this new task.

The maxout model is a feed-forward architecture that uses maxout unit as activation function [34]. Given an input $x \in \mathbb{R}^d$, the maxout hidden layer implements the function

$$h_i(x) = \max_{j\in[1,k]} z_{ij} \tag{2}$$

where $z_{ij} = x^T W_{\cdot\cdot ij} + b_{ij}$, and $W \in \mathbb{R}^{d\times m\times k}$ and $b \in \mathbb{R}^{m\times k}$ are learned parameters. The forward-propagation process of the maxout network is the same as other feed-forward neural networks except that the activation computation follows Eq. (2). For the back-propagation process during training, the gradient for each maxout neuron is always 1, but only the weights corresponding to the piece with the maximum activation within each group $\{z_{ij}|j = 1,\ldots,k\}$ are updated. The max-pooling operation is a winner-take-all action. The maxout neuron, if well-trained, hopefully select the most useful feature invariant to different intra-class modalities, and, at the same time, maximize inter-class variance supervised by the softmax identification signal.

Identification and verification are two supervised signals that are commonly used in deep face recognition [6]. The face identification signal, which is achieved by K-way softmax layer, classifies each image into one of the K candidate identified as follows.

$$L_I = -\sum_{i=1}^{K} y_i \log(p_i) \tag{3}$$

where y_i is the (weak) label, and p_i is the prediction probability distribution over K classes. The verification signal encourages the extracted feature from the same class to be similar [6]. The common loss function is

$$L_V = \begin{cases} \frac{1}{2}\,\| f_i - f_j \|^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max(0,\, m - \| f_i - f_j \|)^2 & \text{if } y_{ij} = 0 \end{cases} \tag{4}$$

where f_i and f_j are the deep feature vectors extracted from a pair of images. $y_{ij} = 1$ means that f_i and f_j are from the same identity, where the L2 distance between the deep feature vectors are minimized. $y_{ij} = 0$ means different identity, where the distance between deep feature vectors is required to larger than a margin m. As in DeepID2 [6], DCMN utilizes a weighted sum to combine these two signals as follows.

$$\min L_I(i) + L_I(j) + \lambda L_V(i, j) \tag{5}$$

where λ is hyper-parameter.

5.2. DCMN architecture

The input to DCMN network is the RGB facial image of a size of 104×96 pixels. Inspired by recent progress on very deep network [35], DCMN contains 9 layers, which are notably deeper than commonly used models, such as Deepface and DeepID. One major difference is that the penultimate layer is replaced by maxout layer. The aim is (1) to learn robust and discriminative features by preserving negative responses, (2) to increase network multiplicity by resembling individual networks. Each of the 7 convolutional layers is followed by a non-linearities such as ReLU. Our model applies filters with a very small 3×3 receptive field to all convolutional layers, which learn to discriminate fine-grained facial structures. Then there is a Dropout [36] layer with a 0.3 dropout-ratio followed Conv7. Spatial pooling is carried out by three max-pooling layers, which follows some of the convolution layers. The last layer is a joint K-way softmax layer and verification supervisory signal is added to the maxout layer to learn a descriptor discriminative to different identities while consistent for images of the same subject. The maxout layer output which contains 1500 neural units is taken as descriptor of input image. The structure of DCMN is shown in Fig. 7.

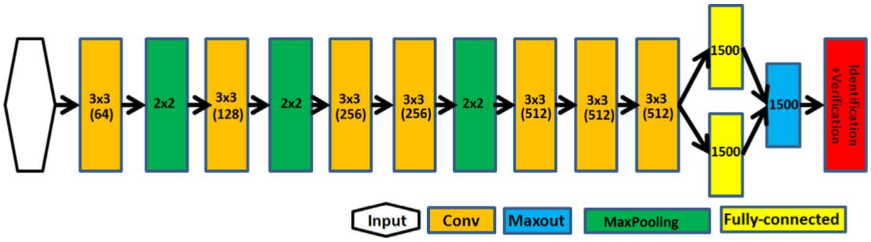


Fig. 7. DCMN Architecture used in our experiment.

It should be noted that the idea of using maxout unit for face verification has been studied in [37]. The difference between our DCMN and the maxout CNN in [37] comes from both the network architecture and the usage of the maxout unit. Our DCMN applies small-size kernels and deeper layers in the network, similar to the VGG architecture. Maxout CNN uses large-size kernels and less layers, similar to the Deepface architecture. Our DCMN applies the maxout unit only in a fully connection layer, but Maxout CNN applies in both the convolution layers and fully connection layers. In the following experiment, we will compared the performance of these two maxout networks.

5.3. Training details

Our DCMN network is trained using SGD(stochastic gradient descent) with a mini-batches of 128 samples (64 pairs) with a momentum of 0.9. To regularize our network, we apply popular weight decay and set it to 0.0005. The dropout layer can also insist to regularize our model with a rate of 0.3 outlined above. Then the hyper-parameter margin m and λ are set to 1 and 0.01, respectively. We set our learning rate to 0.01 at initial stage and then decrease it by factor of 10 when the validation accuracy stops increasing. Overall, the learning rate decreased four times throughout the learning process.

The initialization of the network weights is very important, since traditional initialization procedure of random sampling from a Gaussian distribution can hardly guide the model learning due to instability of training examples and deep net. So we adopted the "xavier" [38] initialization method. Considering the symmetry of human face, the input was randomly sampled and mirrored with 50% probability during training.

6. Experimental results

As the de facto standard on the unconstrained face recognition, LFW has largely promoted the research on feature descriptor, metric learning [13], and deep learning. We examine some well-established methods, and study the comparative performance between LFW and the proposed FGLFW benchmark. This comparison helps us understand how difficult fine-grained face verification is and how should we work towards solving this task? Additional exploration on the human-algorithm cooperative verification is also interesting for future research.

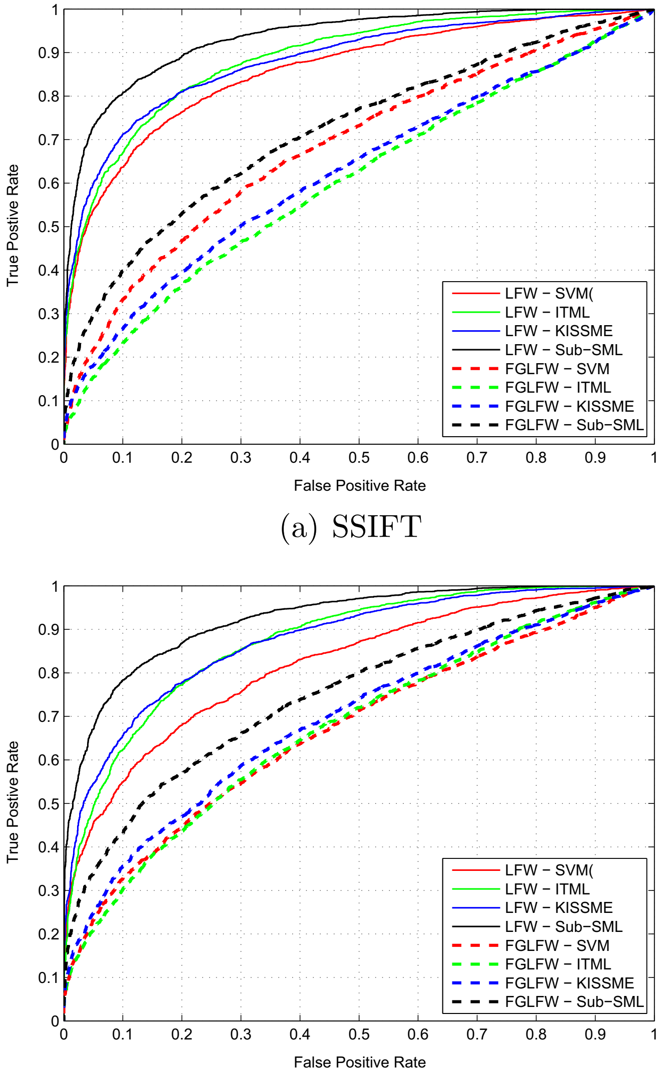
6.1. Comparison on metric learning

The first experiment evaluates the metric learning approaches designed for unconstrained face verification and successfully tested on the LFW. The common objective of metric leaning is to learn a good distance function to reduce the distance of positive pairs and enlarge the distance of negative pairs at the same time. Specifically, we test information theoretic metric learning(ITML) [39], Keep It Simple and Straightforward Metric Learning (KISSME) [40], Sub-SML [41], Support Vector Machine (SVM) [42], under image-restricted protocol,[1] in which only the binary label of face pairs are available.

Two well-known handcrafted face descriptors, i.e. Local Binary Patterns(LBP) [43] and Sparse Scale-Invariant Feature Transform(SSIFT) [44], are used for the base of the comparison on the metric learning. Specifically, we use the LFW-a images for extracting 59-bin uniform pattern LBP histogram in each of the (8 × 15) non-overlapping blocks of the facial image. The descriptor encoding of SSIFT is provided by the authors of [44], which extracts 128 dimensional SIFT descriptors at three scales centered on 9 points as SSIFT.

Table 1
Comparison of mean verification accuracy(%) on LFW dataset and FGLFW under image-restricted setting.

Method	LFW	FGLFW
SVM(SSIFT)	79.03 ± 0.94	65.02 ± 1.50
SVM(LBP)	74.90 ± 1.57	63.87 ± 2.51
ITML(SSIFT)	81.55 ± 1.34	59.52 ± 2.40
ITML(LBP)	80.05 ± 1.82	64.48 ± 1.54
KISSME(SSIFT)	81.75 ± 1.92	61.28 ± 2.59
KISSME(LBP)	79.78 ± 2.18	65.43 ± 1.29
Sub-SML(SSIFT)	85.22 ± 1.19	65.83 ± 2.01
Sub-SML(LBP)	83.92 ± 2.04	67.88 ± 2.32



(a) SSIFT

(b) LBP

Fig. 8. The LFW vs. FGLFW ROC curves of various metric learning methods on (a) SSIFT and (b) LBP descriptors. The accuracy losses of all learned metrics are serious when transferring from LFW to FGLFW.

To reduce the dimension of SSIFT and LBP feature, they are projected into a 300-dimensional PCA subspace before metric learning.

Mean verification accuracy and standard deviation of two feature descriptors using four learned metrics under the image-restricted protocol are listed in Table 1, with the corresponding ROC curves shown in Fig. 8. SSIFT feature based Sub-SML method achieves the

[1] The source codes of ITML [39], KISSME [40], Sub-SML [41] are downloaded from the authors' websites. Default parameters in authors's codes for experiments.

best mean accuracy 85.22% on LFW, but its accuracy deteriorates severely to 65.83% on FGLFW. The accuracy is also down sharply for the other three methods on both features. For instances, ITML with SSIFT drops about 22%, and KISSME with SSIFT drops about 20%.

In general, all tested metric learning methods deteriorate seriously when transferring from LFW to FGLFW, which reflects that fine-grained face verification have universal difficulty for conventional metric learning approaches. Inter-personal difference in FGLFW is much smaller than that of LFW while the intra-class is the same. There might be not any linearly separable direction for inter/intra-class difference, conventional metric learning methods easily generate a misleading results. New study on metric learning, especially the nonlinear technique, should be conducted to address this problem,

6.2. Comparison on high-dimensional feature descriptors

Here we examine two well-known face descriptors, namely Fisher Vector faces (FV) [45], high-dimensional LBP (HDLBP) [26], followed by a Joint Bayesian (JB) [46] model to learn a discriminative metric.[2] They apply principal components analysis (PCA) to first reduce this to 400 dimensions, followed by a Joint Bayesian (JB) [46] model to find a discriminative metric.

Mean verification accuracy of two up-to-date face descriptors are compared in Table 2, with the corresponding ROC curves shown in Fig. 9. The two face descriptors perform similarly on LFW, but become substantially different on FGLFW. This effect is visible in Fig. 9. On FGLFW, high-dimensional LBP performed significantly better than Fisher vector. Note that Fisher vector is a distributional descriptor invariant to translation and distortion, but HDLBP is a localized descriptor extracted at 27 facial landmarks and at five scales, which is effectively characterize localized fine-grained details of human face. The fine-grained description of HDLBP is not highlighted when inter-personal difference is apparent (LFW), but become crucial when recognizing the similarly-looking negative face pairs (FGLFW).

6.3. Comparison on deep learning approaches

Deep convolutional neural network trained by massive labeled outside data have reported the best performance for LFW benchmark. Besides the proposed DCMN, we also implemented three well-known DCNNs for comparison, namely Deepface [47], DeepID2 [6], and VGG-Face [48]. The hyperparameters of three networks are set according to the original papers. CASIA-WebFace [16] database, which contains about 10,000 subjects and 500,000 images, is used for the model training of Deepface and DeepID2. The VGG-Face descriptors are extracted using the off-the-shelf CNN model based on the VGG-Very-Deep-16 CNN architecture as described in [48]. The comparative accuracy is enumerated in Table 3, and the corresponding ROC curves are shown in Fig. 10.

In general, the results show that the proposed DCMN performs better than the VGG-Face, followed by the DeepID2 and Deepfaces. The relatively low accuracy of DeepID2 and Deepfaces on FGLFW indicates that large-size convolution kernels that may not suitable to detect the fine-grained difference between similar faces. The proposed DCMN achieves 98.03% on LFW closing to human-level performance, but its accuracy still drops about 7% on FGLFW. Besides the premier accuracy on both databases, the accuracy loss of DCMN from LFW to FGLFW is much less than the other network. DCMN and VGG-Face both use the small filters to detect the difference between similar faces. The premier performance of DCMN on FGLFW may because the maxout layer of DCMN is adaptive to handle multi-modal intra-class variation (face poses) and, at the same time, the verification signal

Table 2
Comparsion of mean verification accuracy(%) on LFW dataset and FGLFW under image-unrestricted setting with label-free outside data.

Method	LFW	FGLFW
Fisher Vector [45]	93.01 ± 1.17	73.52 ± 1.07
HDLBP [26]	92.11 ± 0.92	80.75 ± 2.01

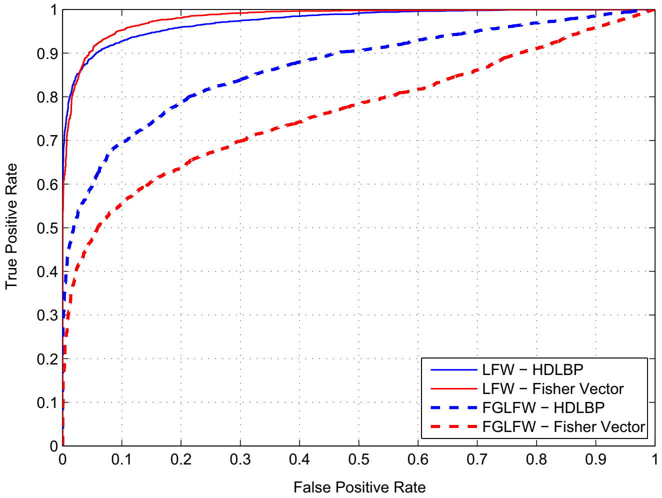


Fig. 9. The comparative LFW vs. FGLFW ROC curves of three state-of-the-art face descriptors. The accuracy loss of Fisher vector is more serious than HDLBP and DCNN when transferring from LFW to FGLFW.

Table 3
Comparison of mean verification accuracy(%) on LFW dataset and FGLFW under image-unrestricted setting using labeled outside data.

Method	# Train	# Net	LFW	FGLFW
DeepFace [47]	0.5 M	1	92.87%	78.78%
DeepID2 [6]	0.5 M	1	95.00%	78.25%
VGG-Face [48]	2.6 M	1	96.70%	85.78%
DCMN	0.5 M	1	**98.03%**	**91.00%**

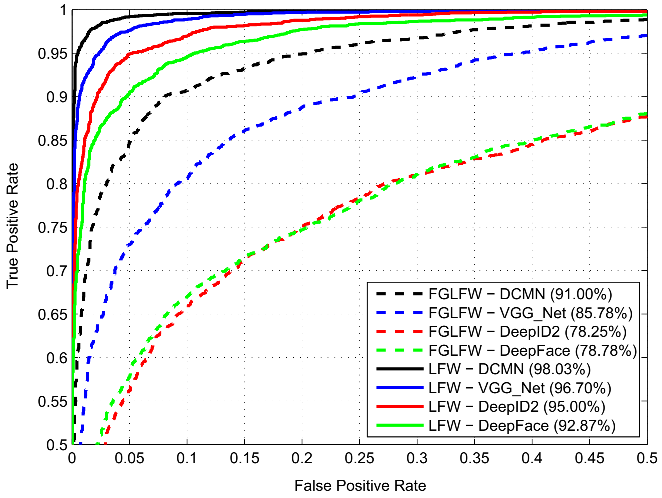


Fig. 10. The comparative LFW vs. FGLFW ROC curves of four deep learning approaches.

[2] The source code of Fisher vector face [45] are downloaded from the authors' websites, Joint Bayesian [46] source code is download from https://github.com/MaoXu.

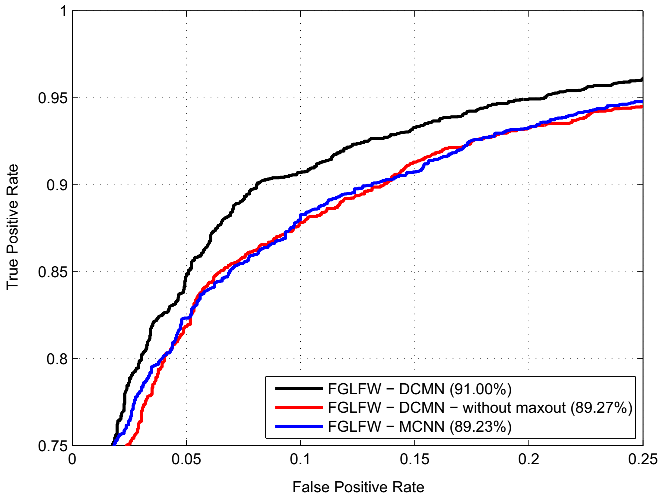


Fig. 11. Contrast experiment of the DCMN and Maxout-CNN [37] on the FGLFW database.

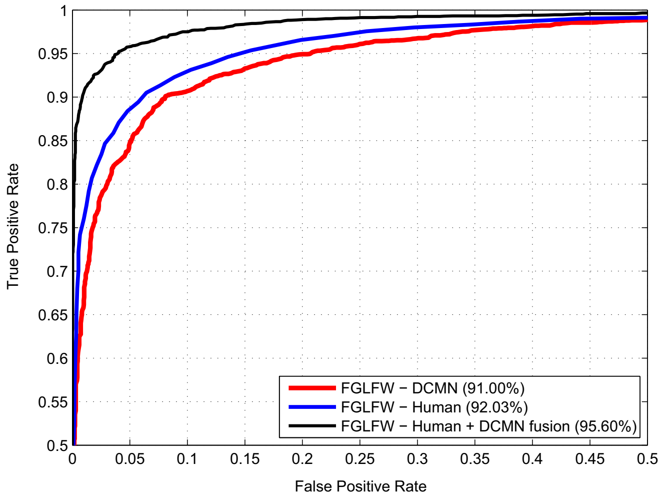


Fig. 12. The fusion performance of the similarity score of human and DCMN on the FGLFW database. There is a strong complementary effect between algorithm and human on this challenging verification task.

reduces the intra-class difference in each modality separately.

We also conducted contrast experiments to investigate the effectiveness of the maxout unit. By replacing the maxout unit in the fully connection layer with the commonly used the RELU unit, we observe that the accuracy drops about 1.7% on the FGLFW experiment, as shown in Fig. 11. We have also conducted contrast experiments that show DCMN outperforms the maxout CNN of [37] by about 1.8%. This suggests that the DCMN network with small-size kernels and deeper layers may be more suitable for capture the fine-grained details. For a fair comparison, we have trained and tested the maxout CNN model in github[3] on our aligned training and test image ensemble. Note that the off-the-shelf feature files offered in the github can achieve similar accuracy to DCMN, but these features are extracted from the authors' own aligned face ensemble [37].

In summary, fine-grained face verification favors localized filters and multi-modality analysis. The accuracy losses of DCMN from LFW to FGLFW, i.e. 7%, is even smaller than human verification, i.e. 8%. Applying deep learning technique to simulate the human fine-grained inspect on localized facial features is an promising way to close the gap between algorithm and human performance.

6.4. *Exploration of human-algorithm partnership*

According to our controlled human survey, there are about 8% deliberate imposters that cannot be detected by human operators. The final experiment explores a important issue on whether algorithms can help human operators to improve fine-grained face verification performance. We fuse the similarity scores of human and DCMN by a simple weighed sum rule. As shown in Fig. 12, although the DCMN performs worse than human operators, a simple score fusion of DCMN and human boosts the accuracy from 92.03% to 95.60%, cutting the error rate in comparison to the human verification by 44.79% [(95.60−92.03)/(100−92.03)=0.4479]. This significant improvement on accuracy suggest somewhat complementary ability on human and algorithms.

It is unclear how human performs fine-grained face verification, but the complementary result indicates human and deep neural network may have different recognition mechanisms. On one hand, developing human-like recognition network may further improve the performance of deep learning. On the other hand, it is very possible to apply more accurate algorithm to help human detect the 8% "successfully cheating" imposters, and thus largely enhance current security of biometric systems. With a sufficient number of "difficult" face pairs and a reliable baseline of human verification, FGLFW would largely facilitate the the study on human-algorithm partnership that may be crucial for real-world application.

7. Summary

We have introduced a novel variant of the well-established LFW database for developing face verification techniques: the Fine-Grained Labeled Face in-the-Wild (FGLFW) collection. The main contributions of the proposed challenge are: First, it provides researchers with a large, challenging database of deliberate imposters from an unconstrained source, with 3000 pairs of human judged similarly-looking faces. Second, our benchmarks focus on fine-grained face similarity, rather than common face discrimination, and test the accuracy of this binary classification based on training with never-before-seen faces. The purpose of this is to gain a more principled understanding of what makes faces different or similar in a fine-grained manner, rather than learn the properties of particular faces. Finally, the benchmarks described in this paper provide a unified testing protocol and an easy means for evaluating the human verification performance and measuring the effectiveness of Human-Algorithm Partnership.

We also tested the validity of our database by evaluating human performance, as well as reporting baseline performance achieved by using state-of-the-art face descriptors, metric learning and deep learning approaches. Empirical results suggest that the FGLFW database indeed provides new challenge current techniques. While humans achieve 92% accuracy on our database, our proposed DCMN yields around 91% success. Finally, we attempted to fuse the similarity score from human and DCMN, and showed human-algorithm fusion *cuts the error rate by a half* in comparison to the human verification. This complementary ability clearly suggests that it is promising to study the algorithms that enhance human's ability to recognize the deliberate imposter.

Conflict of interest

None declared.

[3] https://github.com/AlfredXiangWu/face_verification_experiment

Acknowledgements

This work was partially sponsored by supported by the NSFC (National Natural Science Foundation of China) under Grant nos. 61375031, 61573068, 61471048, and 61273217, the Fundamental Research Funds for the Central Universities under Grant no. 2014ZD03-01, This work was also supported by the Program for New Century Excellent Talents in University, Beijing Nova Program under Grant no. Z161100004916088.

References

[1] W. Deng, J. Hu, J. Lu, J. Guo, Transform-invariant pca: a unified approach to fully automatic facealignment, representation, and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (6) (2014) 1275–1284.

[2] W. Deng, J. Hu, X. Zhou, J. Guo, Equidistant prototypes embedding for single sample based face recognition with generic learning and incremental learning, Pattern Recognit. 47 (12) (2014) 3738–3749.

[3] W. Deng, J. Hu, J. Guo, Extended src: undersampled face recognition via intraclass variant dictionary, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1864–1870.

[4] W. Deng, J. Hu, J. Guo, In defense of sparsity based face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 399–406.

[5] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, CVPR, vol. abs/1503.03832, 2015. Available online: ⟨http://arxiv.org/abs/1503.0383203832⟩.

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

[7] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: a Database for Studying Face Recognition in Unconstrained Environments, 2007.

[8] N. Zhang, W. Deng, Fine-grained lfw database, in: Proceedings of the 9th IAPR International Conference on Biometrics (ICB), 2016.

[9] A.D. Ferencz, E.G. Learned-Miller, J. Malik, Learning hyper-features for visual identification, in: Advances in Neural Information Processing Systems, 2004, pp. 425–432.

[10] A. Ferencz, E.G. Learned-Miller, J. Malik, Building a classification cascade for visual identification from one example, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, ICCV 2005, IEEE, vol. 1, 2005, pp. 286–293.

[11] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D.A. Forsyth, Names and faces in the news, in: 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, IEEE, vol. 2, 2004, pp. II–848.

[12] T.L. Berg, A.C. Berg, J. Edwards, D.A. Forsyth, Whos in the picture, Adv. Neural Inf. Process. Syst. 17 (2005) 137–144.

[13] E. Learned-Miller, G. Huang, A. RoyChowdhury, H. Li, G. Hua, G.B. Huang, Labeled Faces in the Wild: a Survey.

[14] L. Best-Rowden, H. Han, C. Otto, B.F. Klare, A.K. Jain, Unconstrained face recognition: identifying a person of interest from a media collection, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2144–2157.

[15] H.-W. Ng, S. Winkler, A data-driven approach to cleaning large face datasets, in: Proceedings of the 2014 IEEE International Conference onImage Processing (ICIP), IEEE, 2014, pp. 343–347.

[16] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning Face Representation from Scratch, 2014, arXiv preprint arXiv:1411.7923.

[17] D. Miller, I. Kemelmacher-Shlizerman, S.M. Seitz, Megaface: a Million Faces for Recognition at Scale, 2015, arXiv preprint arXiv:1505.02108.

[18] P.J. Phillips, P.J. Flynn, K.W. Bowyer, R.W.V. Bruegge, P.J. Grother, G.W. Quinn, M. Pruitt, Distinguishing identical twins by face recognition, in: Proceedings of the 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011), IEEE, 2011, pp. 185–192.

[19] Z. Sun, A.A. Paulino, J. Feng, Z. Chai, T. Tan, A.K. Jain, A study of multibiometric traits of identical twins, in: SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, 2010, pp. 76670T–76670T.

[20] J. Li, L. Zhang, D. Guo, S. Zhuo, T. Sim, Audio-visual twins database, in: 2015 International Conference on Biometrics (ICB), IEEE, 2015, pp. 493–500.

[21] J. Hu, J. Lu, Y.-P. Tan, Fine-grained face verification: dataset and baseline results, in: 2015 International Conference on Biometrics (ICB). IEEE, 2015, pp. 79–84.

[22] J. Deng, J. Krause, L. Fei-Fei, Fine-grained crowdsourcing for fine-grained recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 580–587.

[23] E. Gavves, B. Fernando, C.G. Snoek, A.W. Smeulders, and T. Tuytelaars, Fine-grained categorization by alignments, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1713–1720.

[24] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based r-cnns for fine-grained category detection, in: European Conference on Computer Vision, Springer, 2014, pp. 834–849.

[25] T. Berg P. Belhumeur, Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 955–962.

[26] D. Chen, X. Cao, F. Wen, J. Sun, Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3025–3032.

[27] S. Huang, Z. Xu, D. Tao, Y. Zhang, Part-stacked cnn for fine-grained visual categorization, 2015, arXiv preprint arXiv:1512.08086.

[28] X. Zhang, H. Xiong, W. Zhou, W. Lin, Q. Tian, Picking deep filter responses for fine-grained image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1134–1142.

[29] S. Reed, Z. Akata, B. Schiele, H. Lee, Learning Deep Representations of Fine-grained Visual Descriptions, 2016, arXiv preprint arXiv:1605.05395.

[30] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear cnn models for fine-grained visual recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1449–1457.

[31] G. Huang, E. Learned-Miller, Labeled Faces in the Wild: Updates and New Reporting Procedures, 2014.

[32] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and simile classifiers for face verification, in: Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 365–372.

[33] A.J. Toole, P.J. Phillips, F. Jiang, J. Ayyad, N. Penard, H. Abdi, Face recognition algorithms surpass humans matching faces over changes in illumination, IEEE Trans. Pattern Anal. Mach. Intell. 29 (9) (2007) 1642–1646.

[34] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A.C. Courville, Y. Bengio, Maxout networks, ICML 28 (3) (2013) 1319–1327.

[35] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-scale Image Recognition, 2014, arXiv preprint arXiv:1409.1556.

[36] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[37] X. Wu, R. He, Z. Sun, A Lightened cnn for Deep Face Representation, 2015, arXiv preprint arXiv:1511.02683.

[38] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks. in: Aistats, vol. 9, 2010, pp. 249–256.

[39] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: Proceedings of the 24th International Conference On Machine Learning. ACM, 2007, pp. 209–216.

[40] M. Koestinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2288–2295.

[41] Q. Cao, Y. Ying, P. Li, Similarity metric learning for face recognition, in: Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, 2013, pp. 2408–2415.

[42] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[43] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.

[44] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? Metric learning approaches for face identification, in: Proceedings of the 2009 12th International Conference on Computer Vision, IEEE, 2009, pp. 498–505.

[45] K. Simonyan, O.M. Parkhi, A. Vedaldi, A. Zisserman, Fisher Vector Faces in the Wild, in: British Machine Vision Conference, 2013.

[46] D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: a joint formulation, in: Computer Vision–ECCV 2012. Springer, 2012, pp. 566–579.

[47] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the 2014 IEEE Conference onComputer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 1701–1708.

[48] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in British Machine Vision Conference, 2015.

Weihong Deng received the B.E. degree in information engineering and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004 and 2009, respectively. From Oct. 2007 to Dec. 2008, he was a postgraduate exchange student in the School of Information Technologies, University of Sydney, Australia, under the support of the China Scholarship Council. He is currently an associate professor in School of Information and Telecommunications Engineering, BUPT. His research interests include statistical pattern recognition and computer vision, with a particular emphasis in face recognition.

Jiani Hu received the B.E. degree in telecommunication engineering from China University of Geosciences in 2003, and the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2008. She is currently a lecturer in School of Information and Telecommunications Engineering, BUPT. Her research interests include information retrieval, statistical pattern recognition and computer vision.

Nanhai Zhang received the B.E. degree in South China University of Technology in 2014. Currently, he is a post-graduate student major in Information and Telecommunications Engineering in BUPT. His research interests include pose-invariant face recognition and deep learning.

Binghui Chen received the B.E. degree in telecommunication engineering from the Beijing University of Posts and Telecommunications (BUPT) in 2015. Currently, he is a post-graduate student major in Information and Telecommunications Engineering. His research interests include pose-invariant face recognition and deep learning.

Jun Guo received B.E. and M.E. degrees from BUPT, China in 1982 and 1985, respectively, Ph.D. degree from the Tohuku-Gakuin University, Japan in 1993. At present he is a professor and the dean of School of Information and Communication Engineering, BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and network management. He has published over 200 papers, some of them are on world-wide famous journals or conferences including SCIENCE, IEEE Trans. on PAMI, IEICE, ICPR, ICCV, SIGIR, etc. His book "Network management" was awarded by the government of Beijing city as a finest textbook for higher education in 2004. His team got a number of prices in national and international academic competitions including: the first place in a national test of handwritten Chinese character recognition 1995, the first place in a national test of face detection 2004, the first place in a national test of text classification 2004, the first place of paper design competition held by IEEE Industry Application Society 2005, the second place in the competition of CSIDC held by IEEE Computer Society 2006.