A Joint Local and Global Deep Metric Learning Method for Caricature Recognition

Wenbin Li[1], Jing Huo[1], Yinghuan Shi[1], Yang Gao[1(✉)], Lei Wang[2], and Jiebo Luo[3]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
liwenbin.nju@gmail.com, {huojing,syh,gaoy}@nju.edu.cn
[2] School of Computing and Information Technology, University of Wollongong, Wollongong, Australia
leiw@uow.edu.au
[3] Department of Computer Science, University of Rochester, Rochester, USA
jluo@cs.rochester.edu

Abstract. Caricature recognition is a novel, interesting, yet challenging problem. Due to the exaggeration and distortion, there is a large cross-modal gap between photographs and caricatures, making it nontrivial to match the features of photographs and caricatures. To address the problem, a joint local and global metric learning method (LGDML) is proposed. First, joint local and global feature representation is learnt with convolutional neural networks to find both discriminant features of local facial parts and global distinctive features of the whole face. Next, in order to fuse the local and global similarities of features, a unified feature representation and similarity measure learning framework is proposed. Various methods are evaluated on the caricature recognition task. We have verified that both local and global features are crucial for caricature recognition. Moreover, experimental results show that, compared with the state-of-the-art methods, LGDML can obtain superior performance in terms of Rank-1 and Rank-10.

Keywords: Caricature recognition · Deep metric learning

1 Introduction

Caricature is a popular artistic drawing style in social media. One caricature is a facial sketch beyond realism, attempting to portray a facial essence by exaggerating some prominent characteristics and oversimplifying the rest. Interestingly, it can be recognized lightly by human at a glance. Moreover, since caricature contains abundant non-verbal information, it is widely used in news and social media. The retrieval between photograph and caricature will be a high demand.

C. V. Jawahar et al. (Eds.): ACCV 2018, LNCS 11364, pp. 240–256, 2019.
https://doi.org/10.1007/978-3-030-20870-7_15

However, there are only a few studies on caricature recognition [1,19,29], which mainly focus on designing and learning mid-level facial attribute features. Moreover, these attributes usually need to be ad-hoc designed and laboriously labeled. Considering the prominent representation ability of deep convolutional neural networks (CNNs), we adopt CNN to learn the features automatically in this paper.

Fig. 1. Local and global similarities between photographs and caricatures.

It is observed, when human verify whether a pair of photograph and caricature belongs to the same person or not, we can first easily connect the special characteristic of photograph with the artistic exaggeration of caricature [26]. For example, the small eyes of Ban Ki moon (Fig. 1(a)), the wing nose of George W. Bush (Fig. 1(b)), the plump lips of Angelina Jolie (Fig. 1(d)), and the pointed chin of Bingbing Fan (Fig. 1(e)). Then, the overall appearance similarity between photograph and caricature from global perspective is taken into consideration [35]. For instance, the long face of Benedict Cumberbatch (Fig. 1(c)).

The above observations imply that the fusion of local and global similarities will benefit measuring the similarity between photograph and caricature. To obtain the fusion of local and global similarities, we present a novel deep metric learning to jointly train a global sub-network and four local part sub-networks. In this method, feature representation and similarity measure are learnt simultaneously, which is end-to-end. Specifically, the global sub-network is used to extract the global features from the whole face for global similarity measure, and the four local part sub-networks are employed to capture the local features from four local parts (*i.e.*, eye, nose, mouth and chin parts) for local similarity measure. By integrating the local and global similarities, we can obtain better similarity measure for photograph and caricature. Thus, the proposed method is termed as *Local and Global Deep Metric Learning (LGDML)*.

In summary, our major contributions include:

- **Joint local and global feature representation:** As a new strategy, joint local and global feature representation learning, is developed for the caricature

recognition task. Based on this strategy, discriminative local and global features of photograph and caricature are learnt, leading to better recognition performance.

- **Unified feature representation and similarity measure learning:** To learn the local and global feature representation and similarity measure (or measure fusion) in a unified framework, we design a novel deep metric learning (DML) method and apply it to the caricature recognition task for the first time. The framework allows us to learn feature representation and similarity measure in a consistent fashion. Under the constraint of metric loss, five single siamese networks are trained, where four are for learning local features and one is for learning global features.
- **Promising results:** Through various experiments, the proposed DML method and the strategy of fusing local and global features prove the most effective for the caricature recognition task. Compared with various network structures, the five single siamese network structures prove the best.
- **Interesting insights:** We verify that an intermediate domain indeed can help reduce the huge semantic gap between two domains when performing a cross-modal recognition task. Moreover, learning feature and metrics simultaneously is more effective for deriving better feature and better metrics than the two-stage process in shallow metric learning.

2 Related Work

2.1 Caricature Recognition

Although many works are proposed for caricature generation [3–5,36,40], there are only few works about caricature recognition [1,19,29]. Klare *et al.* [19] proposed a semi-automatic caricature recognition method by utilizing crowdsourcing. Through crowdsourcing, they define and collect a set of qualitative facial attributes. However, these facial attributes need to be annotated manually, which is difficult and subjective in practical use. On the contrary, Ouyang *et al.* [29] employed attribute learning methodology to automatically estimate the facial attributes. Similar to the aforementioned two works, Abaci *et al.* [1] defined a set of slightly different facial attributes. They adopted a genetic algorithm to evaluate the importance of each attribute and matched the caricature and photograph. Recently, Huo *et al.* [16,17] collected a large caricature dataset and offered four evaluation protocols.

The above methods mainly focus on extracting mid-level facial attributes and conducting experiments on small-scale datasets (*i.e.*, the total number of pairs is less than 200). Our contribution is to design a novel DML-based method on a much larger dataset (*i.e.*, the total number of pairs is more than 1.5×10^5).

2.2 Deep Metric Learning

Compared with conventional shallow metric learning [8,24,32,39], which mainly focuses on learning linear metrics (*e.g.*, Mahalanobis distance based metrics),

DML can learn better non-linear metrics by using deep networks. Several DML methods have been proposed, which can be roughly classified into three categories: (1) CNN combined with metric loss [7,15,28,38,41]; (2) CNN combined with fully connected (FC) layers [11]; (3) Deep structure metric learning [9,13,14].

In the first kind of DML methods, the network structure usually contains two (three) sub-networks, trained by pairwise loss (triplet loss) which is usually used in metric learning. For example, Yi *et al.* [41] adopted a binomial deviance loss to train a siamese neural network for person re-identification task. Cui *et al.* [7] employed a triplet-based DML method to solve the fine-grained visual categorization problem. Huang *et al.* [15] introduced a position dependent deep metric unit, aiming to learn a similarity metric adaptive to local feature structure. In the second kind of DML methods, the FC layers are taken as the metric learning part, while the loss is still cross-entropy loss. A typical representative is MatchNet proposed by Han *et al.* [11]. In the third kind of DML methods, the structure of metric learning is modelled on deep structure (*i.e.*, multilayer perceptron (MLP)) to learn a set of hierarchical nonlinear transformations. However, the inputs of these methods are still hand-crafted features or pre-extracted deep features. Representative works are series of works of Hu and Lu *et al.* [9,13,14].

Our proposed LGDML method belongs to the first category, but the differences include (1) LGDML is a joint local and global multi-view metric method, (2) LGDML focuses on cross-modal verification based on single siamese network and much more sub-networks (*i.e.*, five single siamese networks) are learnt at the same time.

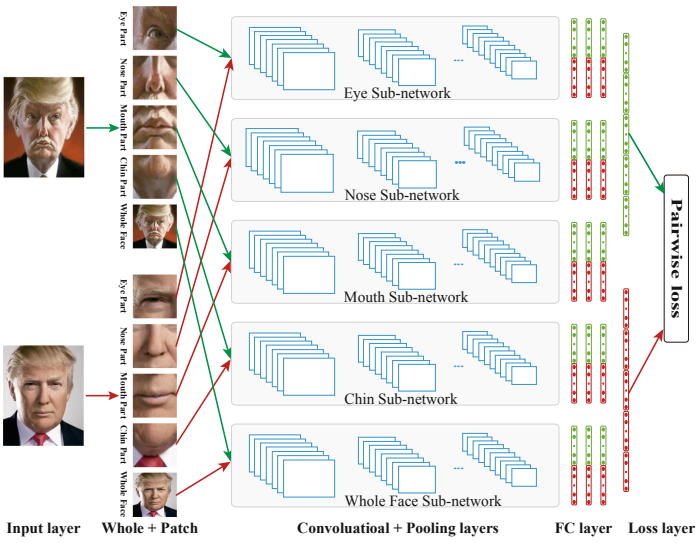


Fig. 2. The framework of the proposed LGDML, containing five single siamese subnetworks.

3 Joint Local and Global Deep Metric Learning

3.1 Network Structure

The framework of LGDML is illustrated in Fig. 2. For each input photograph (caricature), four key parts, *i.e.*, eye, nose, mouth and chin parts, which have abundant local information for recognition (see Fig. 1), are picked and cropped. Combined with the original whole face, these parts are fed into five single sub-networks. In the loss layer, all features of the last FC layers (*i.e.*, Fc8) in these five sub-networks are concatenated. Typically, pairwise loss is adopted to calculate the loss between photograph and caricature. When performing back propagation, the gradients are used for parameter updating of all the sub-networks.

In fact, there should be a total of ten separate sub-networks in this structure for there are ten inputs (*i.e.*, five parts of photograph and five parts of caricature), but it is too difficult and bloated to train this network (*e.g.,* memory limit issue). In order to train this network efficiently, we employ five single siamese sub-networks instead of ten separate sub-networks. Specifically, photograph and caricature share one single sub-network in the same part (*e.g.*, eye part). In other words, two inputs are entered into a single sub-network simultaneously instead of two separate sub-networks which share the same parameters. In addition, compared with traditional siamese network with two identical separate sub-networks or two-tower network with two different separate sub-networks, the single siamese network with only one sub-network can learn better modality invariant features, because data of two modalities are both used to update the same sub-network.

Hence, the advantages of the proposed network structure are that, on one hand, it can leverage the local and global similarities between photograph and caricature simultaneously; on the other hand it can learn good modality invariant features.

3.2 Pairwise Loss Function

For each pair of photograph and caricature, four local metrics and one global metric are learnt together, which can be seen as a multi-view metric. To learn a joint, overall metric, a uniform pairwise loss is used to train all the sub-networks. The goal is to make the fused distance metric between the same-class (*i.e.,* same-individual) pair small and the different-class pair large. From the perspective of different types of metric function, two typical loss functions: Binomial deviance loss [10,41] which focuses on similarity measure and Generalized logistic loss [13, 27] which focuses on distance measure are employed. We describe them in detail as follows:

Binomial deviance loss: Inspired by Yi *et al.* [41], we use cosine similarity to calculate the similarity between two samples, and then adopt binomial deviance to train the network. Given a pair of samples $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d$, and the corresponding

similarity label $l_{ij} \in \{1, -1\}$ (*i.e.*, $l_{ij} = 1$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same class, and $l_{ij} = -1$ otherwise), the formulation can be denoted as follow,

$$\mathcal{L}_{dev} = \ln \left[\exp \Big(-2 \cos(\boldsymbol{x}_i, \boldsymbol{x}_j) l_{ij} \Big) + 1 \right], \tag{1}$$

where $\cos(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denotes the cosine similarity between two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. If $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are from the same class, and the cosine similarity is small, then there will be a large loss of Eq. (1). Otherwise, there will be a small loss of Eq. (1). In this way, the similarity between same-class pair is increased, and the similarity between different-class pair is decreased.

Generalized logistic loss: In metric learning, the major goal is to learn a feature transformation to make the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the transformed space smaller than $\tau - 1$ when $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same class (*i.e.*, $l_{ij} = 1$), and larger than $\tau + 1$ otherwise (*i.e.*, $l_{ij} = -1$). The constraints can be formulated as follow,

$$\begin{aligned} d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) &\leq \tau - 1, l_{ij} = 1 \\ d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) &\geq \tau + 1, l_{ij} = -1, \end{aligned} \tag{2}$$

where $d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$, and $\tau > 1$. For simplicity, the constraints can be written as $l_{ij}\big(\tau - d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)\big) \geq 1$. With the generalized logistic loss function, the loss function is given by

$$\mathcal{L}_{log} = g\Big(1 - l_{ij}\Big(\tau - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\Big)\Big), \tag{3}$$

where $g(z) = \frac{1}{\beta} \log\big(1 + \exp(\beta z)\big)$ is the generalized logistic loss function and β is the *sharpness* parameter.

3.3 Implementation

As AlexNet [21] is a popular and effective network, we take it as the base network in our LGDML. Another reason is that the number of caricature data is still too limited to train deeper networks well, such as VGG-VD [33], GoogLeNet [34] and ResNet [12] *etc.* Usually, the pre-trained AlexNet, which has been trained on the ImageNet dataset, shall be employed. Nevertheless, we observed that directly fine-tuning the pre-trained AlexNet does not produce desirable recognition performance. The reason is that there is a significant semantic gap between the source data (*i.e.*, natural image) and target data (*i.e.*, caricature). To this end, we first adopt other available face image dataset (*e.g.*, PubFig [22]) to fine-tune this pre-trained AlexNet. Afterwards, the fine-tuned AlexNet will be fine-tuned again by caricature data.

During training, we minimize the pairwise loss by performing mini-batch stochastic gradient descent (SGD) over a training set of n photograph-caricature pairs with a batch size of 256 (*i.e.*, 128 pairs). Specifically, we maintain a dropout layer after each FC layer except Fc8 layer, and set the values of momentum and

weight decay to 0.9 and 5×10^{-4} respectively. The filter size of the last FC layer is set to $1 \times 1 \times 4096 \times 4096$, the weights are randomly initialized from a zero-mean Gaussian distribution with 10^{-2} standard deviation, and the biases are initialized to zero. We generate a set of $N = 40$ (*i.e.,* the number of epoches) logarithmically equally spaced points between $10^{-2.7}$ and 10^{-4} as the learning rates.

During forward propagation, a pair of photograph and caricature images are cropped into four pairs of local patches. Then the five pairs of patches (combined with the pair of original images) subtracted their corresponding mean RGB values respectively are fed into five single siamese networks. For each modality, one global feature and four local features can be extracted from the last FC layer. In the final loss layer, the global and local features of each modality are concatenated together to calculate the loss according to the designed cost function. Note that a ℓ_2 normalization layer is added before the loss layer. During back propagation, the parameters of the network are fine-tuned by freezing the first m layers. The reason is that the first several layers mainly learn generic features of images which are transferable between these two modalities [42].

4 Experiments

In this section, we implement various deep networks by changing the structure and loss function. Then, we compare the performance of these methods by conducting caricature recognition task on the WebCaricature dataset [17]. Our implementations are based on the publicly available MATLAB toolbox MatConvNet [37] on one NVIDIA K80 GPU.

4.1 Dataset

PubFig Dataset: To reduce the semantic gap between natural images and caricature images, we choose the PubFig [22] dataset to fine-tune the pre-trained AlexNet. PubFig dataset is a large, real-world face dataset, consisting of a development set and an evaluation set. In our setting, these two subsets are integrated together (36604 images of 200 individuals). After data augmentation, all images (*i.e.*, 512456 images) of the 200 individuals are used to fine-tune a 200-class classification network (*i.e.*, the pre-trained AlexNet). The fine-tuned AlexNet model is named as AlexNet-PubFig.

Caricature Dataset: Our experiments are mainly developed on the WebCaricature dataset, which contains 6042 caricatures and 5974 photographs of 252 individuals. In our experiments, the dataset is divided into two parts, one for training (*i.e.*, 126 individuals) and the other for testing (*i.e.*, the rest 126 individuals). These two parts are disjoint by individual, that is, no individual will appear in both the training and testing sets. Because there are 51 overlapped individuals between PubFig dataset and WebCaricature dataset, the overlapped individuals are only divided into the training set. Besides, in the training set, 30% images of each individual are randomly picked for validation and the rest is used for training.

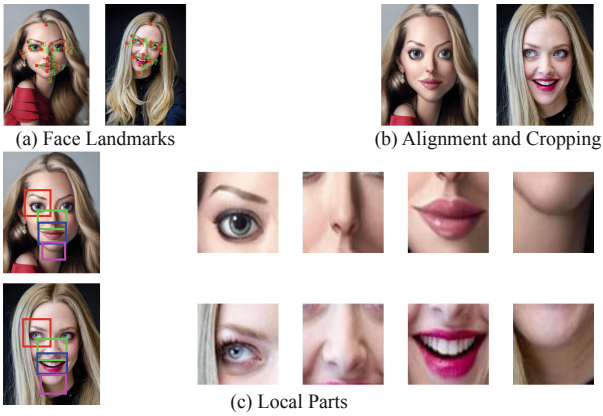


Fig. 3. Illustration of data preprocessing. (a) shows the 17 facial landmarks; (b) exhibits the cropped face images after alignment and rotating; (c) illustrates the cropped local parts.

4.2 Data Preprocessing

Preprocessing: As for each image, 17 landmarks have been provided (Fig. 3(a)) [17]. According to the landmarks, the following face alignment process are employed which includes three major steps: First, each image is aligned by image rotation to make two eyes in a horizontal line. Second, the image is resized to guarantee the distance between two eyes of 75 pixels. Third, the image is cropped by enlarging the bounding box encircled by the face landmarks {# 1, 2, 3, 4} with a scale of 1.2 in both width and height. Finally, the image is eventually resized to 256×320. All the processes are illustrated in Fig. 3.

Augmentation: To better fine-tune our LGDML, we augment the caricature dataset by image flipping in horizontal direction. In this way, we can construct a large-scale image pairs with a magnitude greater than 1.5×10^5. Before using the pre-trained AlexNet, we need to fine-tune this network by utilizing other natural face dataset. In this setting, we also need data augmentation. This time, besides image flipping we also perform random translation inspired by [2]. For each image, we crop a central region 227×227 and randomly sample another 5 images around the image center. Moreover, every image is also horizontally flipped. Thus, 14 images including the resized original image can be obtained after augmentation.

Cropping: To capture the local features of a face, we pick four key parts on the face, *i.e.*, eye part (just left eye), nose part, mouth part and chin part. For the left eye part, landmarks {# 5, 6, 9, 10} (see Fig. 3(a)) are considered, and a rectangle patch is cropped which covers the whole left eye and eyebrow. For the nose part, landmarks {# 9, 10, 11, 12, 13, 14} are taken into account. As for the mouth part, a rectangle patch is cropped according to landmarks {# 13, 14,

15, 16, 17}. So as to the chin part, landmarks {# 3, 15, 16, 17} are considered. Then, all the local patches are resized to 227 × 227 (see Fig. 3(c)).

4.3 Results of Different Deep Network Structures

We report the comparison with different deep methods, which have different network structures. All the methods are evaluated on the caricature recognition task, which is a cross-modal face identification task. Given a caricature (photograph), the goal is to search the corresponding photographs (caricatures) from a photograph (caricature) gallery. For the "Caricature to Photograph" setting, all the caricatures in the testing set (126 individuals) will be used as probes (*i.e.*, 2961 images) and photographs will be used as gallery. Specifically, only one photograph of each individual is selected to the gallery (*i.e.*, 126 images). The setting of "Photograph to Caricature" is similar to the one of "Caricature to Photograph". As these two settings are similar, we only focus on the setting of "Caricature to Photograph". Rank-1 and Rank-10 are chosen as the evaluation criteria.

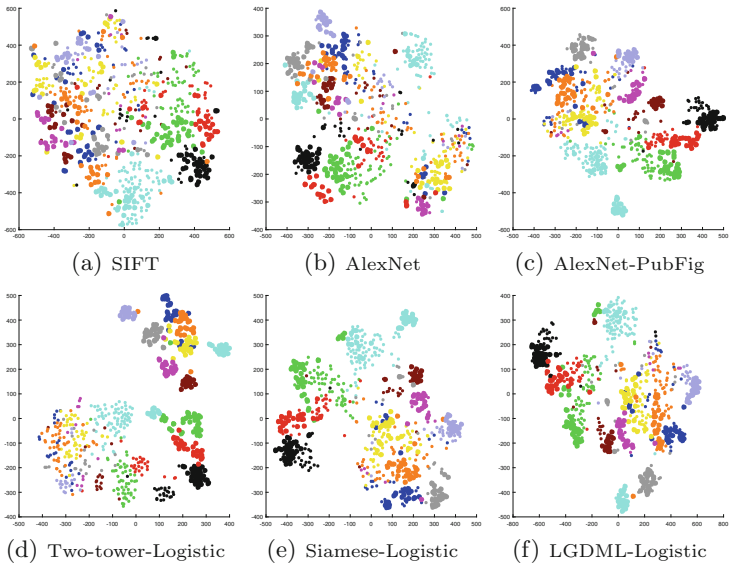
Table 1. Rank-1 (%) and Rank-10 (%) of deep methods with different network structures. Columns 3–4 show the results of raw features. The last two columns exhibit the results after dimensionality reduction by t-SNE.

Structure	Loss	Rank-1	Rank-10	Rank-1 t-SNE	Rank-10 t-SNE
Single	Cross-entropy	24.28	60.79	26.56	54.58
Triplet	Triplet	24.42	61.63	28.57	54.91
Two-tower	Binomial	24.65	62.45	20.63	50.19
Two-tower	Logistic	24.89	62.41	20.42	51.08
Siamese	Binomial	26.21	65.21	30.23	61.06
Siamese	Logistic	27.09	66.60	34.04	62.51
LGDML	Binomial	28.40	**67.65**	36.14	**65.96**
LGDML	Logistic	**29.42**	67.00	**36.27**	64.37

According to the network structure, these deep methods can be divided into five categories as follows:

- **Single Network Methods:** These methods consisting of single network are usually used for classification task. The pre-trained AlexNet-PubFig will be taken as the baseline method without any postprocessing.
- **Siamese Network Methods:** These networks contain two parameter sharing sub-networks which are based on AlexNet-PubFig model. Here, we adopt the single siamese network structure like LGDML. Two loss functions, *i.e.*, binomial deviance loss and generalized logistic loss, would be employed to fine-tune these networks. The depth of back propagation is 11, *i.e.*, updating to conv5 layer.

- **Two-tower Network Methods:** Different from the siamese network, the two sub-networks of two-tower network don't share parameters completely. The binomial deviance loss or generalized logistic loss is used to fine-tune these networks by freezing first several layers (*i.e.*, top 12 layers) which keep the pre-trained parameters unchanged.
- **Triplet Network Methods:** There are three sub-networks with parameter sharing in these networks. Like above networks, these networks also take AlexNet-PubFig as the base network. Moreover, we design a new triplet loss by adding an extra pairwise loss to maximize the use of the provided triplet. Given a triplet $\langle \boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k \rangle$, the new triplet loss can be formalized as $\mathcal{L}_{triplet} = \mu \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 + (1-\mu)(1 + \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 - \|\boldsymbol{x}_i - \boldsymbol{x}_k\|_2^2)_+$, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same class, while $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$ belong to different classes. μ is the hyper-parameter and $(z)_+ = \max(0, z)$ indicates the hinge loss.
- **Our LGDML:** This is the proposed method, containing five single siamese networks. According to the different losses chosen, the proposed method can be named as LGDML-Binomial or LGDML-Logistic.



(a) SIFT (b) AlexNet (c) AlexNet-PubFig

(d) Two-tower-Logistic (e) Siamese-Logistic (f) LGDML-Logistic

Fig. 4. Feature visualization of six representative methods using t-SNE. Different colors denote different individuals (*i.e.*, 11 individuals), big/small dot indicates the photograph/caricature modality, respectively. (Color figure online)

It is worth noting that although we do not explicitly compare the proposed LGDML with other existing cross-modal methods, the competitive network structures implicitly represent some existing methods. For example, in [30], a two-tower network combined with the contrastive loss was employed to solve the

near-infrared heterogeneous face recognition problem. In addition, [31] adopted a triplet loss to train a face recognition network, which is equivalent to the triplet network in our experiments. All these deep methods aim to learn a good feature representation. Hence, for the first four deep methods, a 4096-dimensional feature is extracted from the first FC layer (*i.e.*, Fc6 layer), which is proved to be more expressive than Fc7 and Fc8 in feature representation. LGDML extracts a 20480-dimensional feature by integrating all the features of the four local parts and the whole image. A popular dimensionality reduction method t-SNE [25] is also employed to make all features into a same dimensionality (*i.e.*, 300). Table 1 reports the results of all the methods. LGDML achieves the best rank-1 and rank-10 performance with 29.42% and 67.65%. When performing dimensionality reduction, the results are 36.27% and 65.95%. From the results, we can observe that:

Influence of loss function: Binomial deviance loss (denoted as Binomial) performs similar with generalized logistic loss (denoted as Logistic). While the triplet loss (denoted as Triplet) does not achieve promising results, the reason may be that three separate sub-networks are employed in the triplet network, which cannot learn good modality invariant features.

Influence of network structure: Under the same loss function setting, two-tower structure performs worse than the single siamese structure. The reason is that single siamese structure is more tend to learn modality invariant feature (see Fig. 4(d), (e)). From Fig. 4(f), we can see that the features learnt from LGDML are blended together in the modality, but are distinguishable between different individuals. LGDML can learn both modality invariant and discriminant features, which makes LGDML achieve the state-of-the-art result.

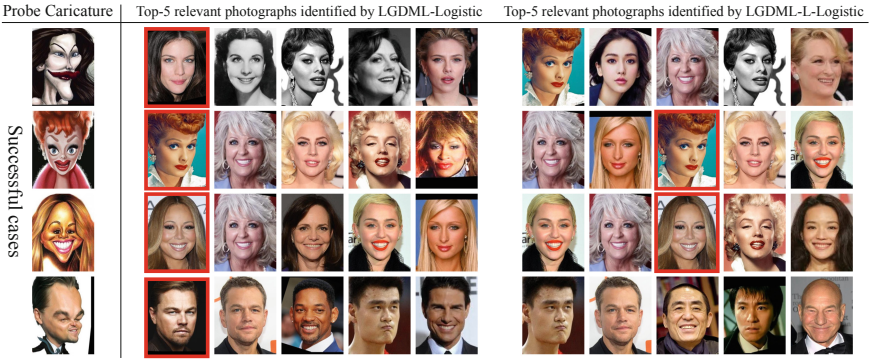


Fig. 5. Success cases of caricature recognition results by LGDML and LGDML-Local. For each probe caricature, top 5 relevant photographs are exhibited, where the photographs annotated with red rectangular boxes are the ground-truth. (Color figure online)

4.4 Local and Global Methods

LGDML can learn local and global features simultaneously. To illustrate the effectiveness of fusion of the local and global features, we reduce LGDML to a simpler variant by only learning local features namely LGDML-Local. It can be seen that if we only learn local features (see Table 2), the result becomes worse due to the lack of global information. We also reduce LGDML to another simpler variant by only learning global features namely LGDML-Global. In fact, LGDML-Global is same as AlexNet-PubFig-Siamese in Table 1. The results in Table 2 show that it is beneficial to integrate local and global features. A clear effect of this integration can also be seen in Fig. 5. We can see that LGDML is obviously superior to LGDML-Local.

Table 2. Local and global methods.

Method	Loss	Type	Rank-1 (%) t-SNE	Rank-10 (%) t-SNE
LGDML-Local	Binomial	Local	23.57	50.35
LGDML-Local	Logistic	Local	21.65	45.80
LGDML-Global	Binomial	Global	30.23	61.06
LGDML-Global	Logistic	Global	34.04	62.51
LGDML	Binomial	Local+Global	36.14	**65.96**
LGDML	Logistic	Local+Global	**36.27**	64.37

4.5 Indirect and Direct Fine-Tuning

From Table 3, we can see that if we directly perform fine-tuning on the AlexNet which is pre-trained on the ImageNet, the rank-1 performance can only reach 18.34% (*i.e.*, the result of AlexNet-Siamese-Logistic). However, if we perform fine-tuning on the AlexNet-PubFig, which is fine-tuned based on the pre-trained AlexNet, the rank-1 performance can reach 34.04% (AlexNet-PubFig-Siamese-Logistic). This inspires us that when we perform fine-tuning on two domains that have huge semantic gap (*i.e.*, natural image and caricature), we can resort to an intermediate domain (*i.e.*, natural face image) between these two domains first.

4.6 Deep and Hand-Crafted Features

In addition to deep features, we also compare deep methods with hand-crafted feature extraction methods. Three hand-crafted features will be extracted for each image respectively, that is, LBP, Gabor and SIFT [1,19,29]. For LBP feature, the original image (256×320) is partitioned into 4×5 patches of 64×64.

Table 3. Indirect and direct fine-tuning.

Base network	Architecture	Loss	Rank-1 t-SNE	Rank-10 t-SNE
AlexNet	Siamese	Binomial	17.76	39.28
AlexNet	Siamese	Logistic	18.34	40.19
AlexNet-Pubfig	Siamese	Binomial	30.23	61.06
AlexNet-Pubfig	Siamese	Logistic	**34.04**	**62.51**

Table 4. Deep and hand-crafted features.

Base network	Feature/loss	Rank-1 t-SNE	Rank-10 t-SNE
–	LBP	1.65	12.23
	Gabor	3.24	15.30
	SIFT	9.56	29.08
AlexNet	Cross-entropy	14.39	36.68
AlexNet-Pubfig	Cross-entropy	**26.56**	**54.58**

In each patch, a 30-dimensional uniform LBP feature is extracted. We can get a 600-dimensional LBP feature after combining the features of all patches. To extract Gabor feature, the original 256×320 image is resized to 256×256 and 40 filters are used. After filtering, the filtered image is down sampled to $\frac{1}{16}$ of its original size. Then, the vectorized images are concatenated to obtain a 10240-dimensional Gabor feature. For SIFT feature, the original image is divided into 10×13 patches of 64×64 with a stride of 20 pixels. In each 64×64 patch, a 32-dimensional SIFT feature is extracted. Then all the features are concatenated to get a 4160-dimensional SIFT feature.

Hand-crafted features perform poorly on this task (see Table 4), which reflects the difficulty of this task. Interestingly, the pre-trained AlexNet achieves better performance than the best hand-crafted feature (*i.e.*, SIFT), although the feature of AlexNet is just learnt from natural images. AlexNet-PubFig, which is just fine-tuned by natural face images, achieves significant performance improvement (more than 15% performance improvement in rank-1). This verifies again, through the caricature recognition task, that, compared with hand-crafted methods, deep learning indeed has stronger ability of feature representation.

4.7 Deep and Shallow Metric Learning

We compare our DML method with traditional shallow metric learning methods. Several state-of-the-art shallow metric learning methods are picked, including large margin nearest neighbor (LMNN) [39], information-theoretic metric learning (ITML) [8], KISSME [20], logdet exact gradient online (LEGO) [18], online algorithm for scalable image similarity (OASIS) [6] and OPML [23]. All these methods learn from the deep features extracted from the AlexNet-PubFig

network. For fair comparison, all features are reduced to features with a suitable dimensions (*i.e.*, 300) by PCA. We summarized the results in Table 5. From the results, we can see that most shallow metric learning methods can hardly improve the performance. Among them, ITML achieves the best result (just about 2% performance improvement in rank-1). In contrast, DML methods can further improve the performance.

The above results can be explained as follows. Traditional shallow metric learning generally focuses on learning new feature representation based on the given input feature representation. It is a two-stage process, in which feature extraction and distance measure are usually separated. The given input feature representation has limited the upper bound of the optimization of metric learning algorithms, and their quality directly affects the performance improvement of metric learning. In other words, metric learning could make large performance improvement on weak feature representation (*e.g.*, hand-crafted features), but can only make a small improvement on powerful feature representation (*e.g.*, deep features). In contrast, DML integrates feature extraction and distance measure together. It can learn feature and metrics simultaneously, and makes them to work best with each other. In this way, DML can achieve better feature and better metrics. In addition, shallow metric learning methods usually learn a linear transformation, which cannot effectively capture the non-linear structure in the data. On the contrary, the non-linear features learnt from DML, *e.g.*, our proposed LGDML, are more capable in this regard.

Table 5. Deep and shallow metric learning.

Method	Rank-1 (%) PCA	Rank-10 (%) PCA
AlexNet-PugFig	23.74	60.15
KissMe	21.28	55.56
OASIS	21.61	64.00
OPML	23.98	61.03
LEGO	24.38	60.22
LMNN	25.60	62.60
ITML	26.02	63.07
Siamese-Logistic	26.98	66.26
LGDML-Binomial	28.06	**66.57**
LGDML-Logistic	**28.88**	66.30

5 Conclusions

Caricature recognition is a challenging and interesting problem, but has not been sufficiently studied. Furthermore, the existing methods mainly pay attention to

mid-level facial attributes, which are expensive to annotate manually, and need ad-hoc settings. In this paper, taking advantage of the strong representation ability of deep learning and discriminative transformation of metric learning, we propose LGDML to solve the caricature recognition task. In LGDML, local and global features of caricature are jointly learnt. In addition, metric loss is chosen to optimize the entire network, allowing feature representation and distance metric to be learnt simultaneously. The experiments have been conducted extensively to evaluate all the comparable methods, and our proposed LGDML outperform all the other methods.

Acknowledgements. This work is supported by the National NSF of China (Nos. 61432008, 61673203, 61806092, U1435214), Primary R&D Plan of Jiangsu Province, China (Nos. BE2015213), Jiangsu Natural Science Foundation (Nos. BK20180326), CCF-Tencent RAGR (Nos. 20180114) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

1. Abaci, B., Akgul, T.: Matching caricatures to photographs. SIVP **9**(1), 295–303 (2015)
2. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR, pp. 3908–3916 (2015)
3. Akleman, E.: Making caricatures with morphing. In: SIGGRAPH, p. 145 (1997)
4. Akleman, E., Reisch, J.: Modeling expressive 3D caricatures. In: SIGGRAPH, p. 61 (2004)
5. Brennan, S.E.: Caricature generator: the dynamic exaggeration of faces by computer. Leonardo **40**(4), 392–400 (2007)
6. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. JMLR **11**, 1109–1135 (2010)
7. Cui, Y., Zhou, F., Lin, Y., Belongie, S.: Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In: CVPR, pp. 1153–1162 (2016)
8. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML, pp. 209–216 (2007)
9. Duan, Y., Lu, J., Feng, J., Zhou, J.: Deep localized metric learning. TCSVT **28**, 2644–2656 (2017)
10. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, Heidelberg (2009). https://doi.org/10.1007/978-0-387-84858-7
11. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: MatchNet: unifying feature and metric learning for patch-based matching. In: CVPR, pp. 3279–3286 (2015)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
13. Hu, J., Lu, J., Tan, Y.P.: Discriminative deep metric learning for face verification in the wild. In: CVPR, pp. 1875–1882 (2014)
14. Hu, J., Lu, J., Tan, Y.P.: Deep metric learning for visual tracking. TCSVT **26**(11), 2056–2068 (2016)
15. Huang, C., Loy, C.C., Tang, X.: Local similarity-aware deep feature embedding. In: NIPS, pp. 1262–1270 (2016)

16. Huo, J., Gao, Y., Shi, Y., Yin, H.: Variation robust cross-modal metric learning for caricature recognition. In: ACMMM Workshop, pp. 340–348. ACM (2017)
17. Huo, J., Li, W., Shi, Y., Gao, Y., Yin, H.: WebCaricature: a benchmark for caricature face recognition. In: BMVC (2018)
18. Jain, P., Kulis, B., Dhillon, I.S., Grauman, K.: Online metric learning and fast similarity search. In: NIPS, pp. 761–768 (2009)
19. Klare, B.F., Bucak, S.S., Jain, A.K., Akgul, T.: Towards automated caricature recognition. In: ICB, pp. 139–146 (2012)
20. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR, pp. 2288–2295 (2012)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
22. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV, pp. 365–372 (2009)
23. Li, W., Gao, Y., Wang, L., Zhou, L., Huo, J., Shi, Y.: OPML: a one-pass closed-form solution for online metric learning. Pattern Recogn. **75**, 302–314 (2018)
24. Li, W., Huo, J., Shi, Y., Gao, Y., Wang, L., Luo, J.: Online deep metric learning. arXiv:1805.05510 (2018)
25. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. JMLR **9**(Nov), 2579–2605 (2008)
26. Mauro, R., Kubovy, M.: Caricature and face recognition. Mem. Cogn. **20**(4), 433–440 (1992)
27. Mignon, A., Jurie, F.: PCCA: a new approach for distance learning from sparse pairwise constraints. In: CVPR, pp. 2666–2672 (2012)
28. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR, pp. 4004–4012 (2016)
29. Ouyang, S., Hospedales, T., Song, Y.-Z., Li, X.: Cross-modal face matching: beyond viewed sketches. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9004, pp. 210–225. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16808-1_15
30. Reale, C., Nasrabadi, N.M., Kwon, H., Chellappa, R.: Seeing the forest from the trees: a holistic approach to near-infrared heterogeneous face recognition. In: CVPR Workshop, pp. 54–62 (2016)
31. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: CVPR, pp. 815–823 (2015)
32. Shi, Y., Li, W., Gao, Y., Cao, L., Shen, D.: Beyond IID: learning to combine non-IID metrics for vision tasks. In: AAAI, pp. 1524–1531 (2017)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
34. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR, pp. 1–9 (2015)
35. Tanaka, J.W., Farah, M.J.: Parts and wholes in face recognition. Q. J. Exp. Psychol. **46**(2), 225–245 (1993)
36. Tseng, C.C., Lien, J.J.J., Member, I.: Colored exaggerative caricature creation using inter-and intra-correlations of feature shapes and positions. IVC **30**(1), 15–25 (2012)
37. Vedaldi, A., Lenc, K.: MatConvNet: convolutional neural networks for MATLAB. In: ACMMM, pp. 689–692 (2015)
38. Wang, J., et al.: Learning fine-grained image similarity with deep ranking. In: CVPR, pp. 1386–1393 (2014)
39. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: NIPS, pp. 1473–1480 (2005)

40. Yang, W., Toyoura, M., Xu, J., Ohnuma, F., Mao, X.: Example-based caricature generation with exaggeration control. TVC **32**(3), 383–392 (2016)
41. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: ICPR, pp. 34–39 (2014)
42. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NIPS, pp. 3320–3328 (2014)