WebCaricature: a benchmark for caricature recognition

Jing Huo[1]
huojing@nju.eu.cn

Wenbin Li[1]
liwenbin.nju@gmail.com

Yinghuan Shi[1]
syh@nju.edu.cn

Yang Gao[1]
gaoy@nju.edu.cn

Hujun Yin[2]
hujun.yin@manchester.ac.uk

[1] State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing, China

[2] School of Electrical and Electronic Engineering
The University of Manchester
Manchester M13 9PL, U.K.

Abstract

Studying caricature recognition is fundamentally important to understanding of face perception. However, little research has been conducted in the computer vision community, largely due to the shortage of suitable datasets. In this paper, a new caricature dataset is built, with the objective to facilitate research in caricature recognition. All the caricatures and face images were collected from the Web. Compared with two existing datasets, this dataset is much more challenging, with a much greater number of available images, artistic styles and larger intra-personal variations. Evaluation protocols are also offered together with their baseline performances on the dataset to allow fair comparisons. Besides, a framework for caricature face recognition is presented to make a thorough analyze of the challenges of caricature recognition. By analyzing the challenges, the goal is to show problems that worth to be further investigated. Additionally, based on the evaluation protocols and the framework, baseline performances of various state-of-the-art algorithms are provided. A conclusion is that there is still a large space for performance improvement and the analyzed problems still need further investigation.

1 Introduction

In the past few years, face recognition performance, even in unconstrained environments, has improved substantially. For instance, the recognition accuracy on the LFW dataset has reached 99%, even outperforming most humans [29]. However, caricature recognition performances by computers are still low [13, 16].

Caricatures are facial drawings by artists with exaggerations of certain facial parts or features. The exaggerations are often beyond realism and yet the caricatures are effortlessly recognizable by humans. In some cases, people even find that caricatures with exaggerations are easier to recognise as compared to face photos [21, 26]. However, this is not the case

Table 1: Summary of existing datasets relating to caricature recognition.

	Dataset information		
	Number of subjects	Number of images	Dataset purpose
Klare et al. [16]	196 subjects	392 images	Caricature recognition
Abaci and Akgul [1]	200 subjects	400 images	Caricature recognition
Crowley et al. [6]	Dev: 1,088 subjects NPG: 188 subjects Train: 496 subjects	Dev: 8,528 images NPG: 3,128 images Train: 257,000 images	Painting retrieval
Mishra [23]	100 subjects	Caricature: 8,928 images Face: 1,000 images	Cartoon recognition
This paper	252 subjects	12,016 images	Caricature recognition

for computers. For computers, the basic goal of caricature recognition consists of deciding whether a caricature image and a photo are from the same person. As caricatures are of varying styles and different shape exaggerations, such recognition is difficult.

Caricatures have long fascinated psychologists, neuroscientists and now computer scientists and engineers in their seemingly and grossly distorted views of veridical faces while still possessing distinctively recognizable features of the subjects [21, 27, 28]. Studying caricatures can offer valuable insights to how face recognition is robustly performed by humans. Studying caricature recognition can also lead to a better understanding of human perception of faces. For example, studies on caricature recognition [21] have shown that faces may be encoded as distinctive features deviated from prototype faces in human brain. It is thus interesting to see what will happen if this result is used in the current deep learning methods to explicitly encode face features as their distinctiveness from the prototypes. However, this has been rarely studied. Therefore, it is fundamentally imperative to study caricature recognition to shed light on the intrinsic nature of human perception of faces. Computer scientists can take insights gained from psychological studies to develop machine learning methods to further improve caricature and face recognition performance.

Currently, there is only limited work on caricature recognition. Besides, there also lack suitable caricature datasets for such research. In fact, there are only four publicly available datasets [1, 6, 16, 23] that are related to caricature recognition, shown in Table 1. The dataset of Klare et al. [16] only has 392 images, while the dataset of Abaci and Akgul [1] has 400 images, which certainly limits the study of caricature recognition. Another limitation of these two datasets is that they do not provide benchmark evaluation protocols, which may prove difficult to compare studies and performances. The dataset of Mishra et al. [23] has more images, but only from 100 subjects. Another related dataset is from Crowley et al. [6]. However, this dataset was developed for painting retrieval rather than caricature recognition.

The main contribution of this paper is that we have built a new, larger dataset of caricatures and photos (called WebCaricature[1]) with over 12,000 images to facilitate the study on caricature recognition. Examples of photos and caricatures of this new dataset are shown in Fig. 1. This dataset is much larger than any existing ones and contains more challenging intra-class variations for both photos and caricatures. In addition, we have provided facial landmark annotations on each images of the entire dataset, as well as several evaluation protocols and their baseline performances for comparison.

On caricature recognition, Klare et al. [16] proposed to use attribute features, labeled by human via Amazon's Mechanical Turk service. Logistic regression (LR) [8], multiple kernel
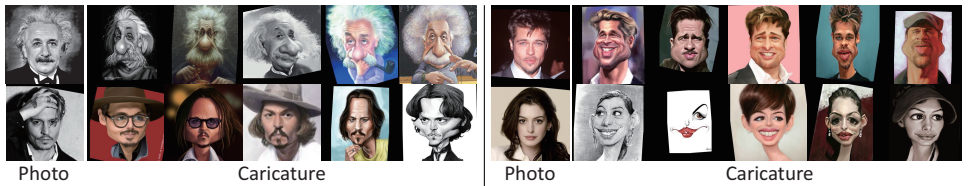
Figure 1: Examples of photos and caricatures. For each person, a photo is given in the first column. The next five columns are corresponding caricatures of various artistic styles, exhibiting large intra-class variability.

learning (MKL) [3], and support vector machines (SVM) [5] were then used to calculate the similarity of a caricature and a photo based on these features. Authors in [24] proposed to learn a facial attribute model. Facial attribute features were then combined with low-level features for recognition using canonical correlation analysis (CCA). Abaci and Akgul [1] proposed a method to extract facial attribute features for photos. For caricatures, the attribute features were manually labeled. Then the weights of these attribute features were learned by a genetic algorithm (GA) or LR and then used for recognition.

The existing work is monotonous with main effort on extracting facial attribute features. Little work [13] has discussed the possibility of applying low-level features and/or deep learning features for caricature recognition. Besides, the existing work did not consider the detection and facial landmark localization of caricatures, an important process for automatic recognition. To complement this absence of related work, a caricature recognition framework is presented. Similar to the current face recognition framework, we present challenges of caricature recognition within the framework as compared to face recognition. Based on this newly constructed dataset and its evaluation protocols, a combination of many state-of-the-art methods under the framework are tested. The experiments not only help to demonstrate the challenges but also offer baseline performances. With the presented framework, future researchers are advocated to focus more on key challenges rather than stalling at attribute feature learning. To summarize, in additional to a new large caricature dataset, Web-Caricature, this paper presents a framework for caricature recognition. With the framework, we analyze the challenges of caricature recognition and hence provide baseline performances on the dataset.

2 Dataset Collection

2.1 Collection Process

For collecting both face photos and caricature drawings, we first drew up a list of names of celebrities. For each person, we manually searched for caricatures and photos and saved these images. All the photos were searched using the Google image search, while caricatures were mainly from the Google image search and the Pinterest website.

After all the images were downloaded, a program was written to detect and remove duplicated caricatures and photos. This was done by extracting features of any two caricatures or photos, pairs of features with large similarities were selected. We then manually checked whether they were indeed duplicated images. The process resulted in a dataset of 252 subjects with a total of 6042 caricatures and 5974 photos. For each person, the number of

caricatures ranges from 1 to 114 and the number of photos from 7 to 59.

2.2   Labeling Process

For each caricature, 17 landmarks were manually labeled. The first four landmarks are the basic face contours. The next four are eye brows. Landmarks 9-12 are eye corners. The 13th landmark is nose tip. The 14th-17th are mouth contours. Details can be found from the dataset website.

The landmark labeling procedure for photos was different. For each photo, firstly we used the facial landmark detection software provided by the Face++ software [22]. The software could locate up to 83 facial landmarks, from which we used only the parts of the face landmarks that corresponded to the 2nd-17th landmarks labelled on the caricatures. The first landmark was then manually labeled for each of the photos. Note that there were some photos the software did not label well or failed to locate landmarks. For those cases, we manually corrected the labeling or manually labeled them with the same scheme as for the caricatures. After the above two processes, 17 landmarks for each caricature and photo were obtained.

3   Evaluation Protocols

Generally, face recognition can be categorized into two categories: verification and identification. As the two kinds of recognition generally deal with different problems. To promote the study of both caricature verification and identification, four experimental protocols covering both scenarios are developed on the constructed WebCaricature dataset. All the protocols are public available along with the dataset.

Caricature Verification. The task of caricature verification is to verify whether a caricature and a photo are from the same person. Therefore, the algorithm is presented with a pair of images (a caricature and a photo) and the output is either yes or no. To evaluate the performance of this protocol, we have built two settings, similar to that used for LFW [10]. One is image restricted setting and the other is image unrestricted setting. For each setting, the dataset is divided into training and testing sets.

In the image restricted setting, there are two views. View 1 is for parameter tuning and View 2 for testing. Pairs of images are provided with the information on whether the pairs are from the same person. There is no extra identity information. For training, only the provided pairs should be used. In the image unrestricted setting, there are also two views. In this setting, training images together with person identities are given. Therefore, as many pairs as possible can be formulated for training. For View 1 of both restricted and unrestricted settings, the proportion of subjects in training and testing were set to close to 9:1. For View 2, 10-fold cross validation should be used. Therefore, to report results on View 2, training must be carried out for ten times; and each time, 9 folds of the data are used for training and the remaining fold for testing. For these two settings, researchers are encouraged to use the receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), VR@FAR=0.1%, and VR@FAR=1% to report performance, where VR@FAR=0.1% corresponds to the verification rate (VR) with false accept rate (FAR) equal to 0.1%, and VR@FAR=1% denotes VR with FAR of 1%.

Caricature Identification. The task of caricature identification can be formulated into two settings. One is to find the corresponding photo(s) of a given caricature from a photo
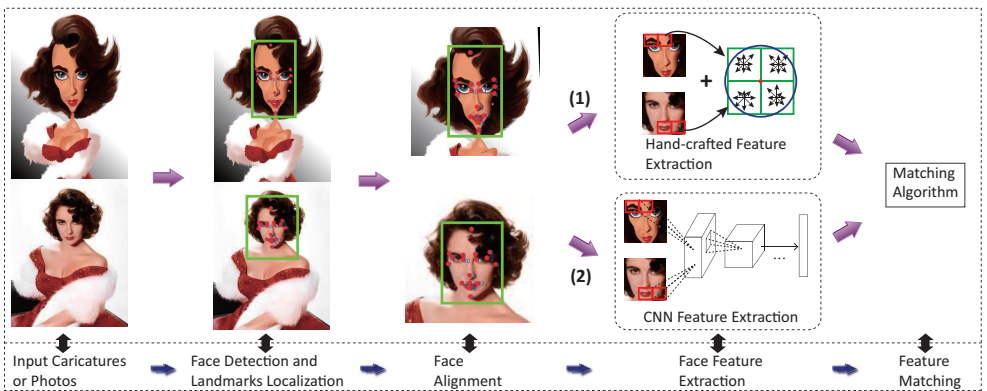
Figure 2: Caricature recognition framework. The first step is face detection and landmark localization. Then faces are aligned. With aligned face images, features are extracted. As illustrated, two kinds of feature extraction methods can be used, hand-crafted or convolutional neural network based. The last step is to measure the similarity of extracted features.

gallery (denoted as Caricature to Photo or C2P). The second is to find the corresponding caricature(s) for a given photo from a set of caricatures (denoted as Photo to Caricature or P2C). For each of the two settings, there are two views. View 1 is for development and parameter tuning and View 2 for testing. While generating data for these two views, subjects were evenly split to training and testing sets. For View 2, the dataset was randomly split for ten times. Thus for results reporting, the algorithm should run ten times and report the average results. Besides, for C2P, photos are used as the gallery and caricatures are used as the probe. For each subject in the gallery set, only one photo is randomly selected. Reversely, for P2C, caricatures are used as the gallery and each subject has only one randomly selected caricature in the gallery. For these two settings, we encourage researchers to use the cumulative match curve (CMC) for evaluation and report average Rank-1 and Rank-10 results.

4 Recognition Framework and Challenges

As reviewed in Section 1, most existing work on caricature recognition tries to extract descriptive attribute features for recognition. In this paper, we illustrate how to solve the recognition problem without using attributes, as they are subjective and require extensive manual labeling. The proposed recognition framework is shown in Fig. 2. There are four main steps in the framework. The process is similar to the traditional face recognition process, albeit more challenges at each step.

Caricature Face Detection and Landmark Localization. The first step of the recognition process is to perform caricature face detection and landmark localization. For the detection, as caricatures can vary greatly in artistic styles and facial appearances, it is more difficult to find detection patterns of caricatures compared to real faces in photos. For localizing landmarks, the shapes of caricatures and the positions of eyes, noses and mouths can be exaggerated and sometimes appear at odd positions beyond realism, making most of the existing localization techniques unusable. Although this paper is devoted mainly for studying caricature recognition, further research on automatic caricature detection and landmark
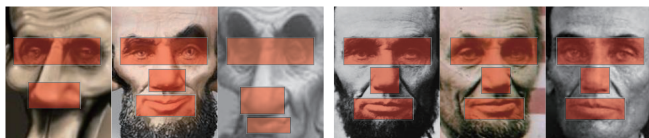
Figure 3: Illustration of the misalignment problem of caricatures. Nose and mouth parts appear at different positions on the aligned caricature images (first three). Photos are fairly well aligned (last three).

localization can also be performed on the proposed dataset, as our dataset only provides 17 landmarks for each caricature.

Face Alignment. The second step is face alignment. For this task, the objective is to make the eyes, noses or mouths appear at the same positions on two cropped face images, such that the two images are comparable. Three alignment methods (details can be found in Section 5.1), which work well for traditional face recognition, were tested in this paper. However, we found that in caricature recognition, caricatures can still be misaligned even after alignment is applied, because many facial parts are exaggerated and distorted. An illustration of misaligned caricatures after applying the eye-based alignment method is given in Fig. 3. The eye-based alignment method forces the eye distance to a fixed number and the center of two eyes at a fixed position in an aligned face image. As can be seen, the photos are well aligned. Eyes of the caricatures are also well aligned. However, there are mismatches in nose and mouth areas of the caricatures. Hence for caricature recognition, more sophisticated methods need to be developed to address the misalignment problem.

Face Feature Extraction. The goal of face feature extraction is to extract robust and discriminant person specific features. Besides the challenges of traditional face feature extraction, such as illumination, expression and viewing angle variations, one major challenge for caricature feature extraction is to remove the variation across different drawing styles (also identified as modality variation, as photos and caricatures can be seen as from two modalities). Besides, if misalignment problem is not addressed well in the previous step, the feature extraction step should also take the exaggeration and distortion problems into consideration, making the feature extraction even more difficult.

In the experiments, we have tested two kinds of feature extraction methods. The first kind is hand-crafted feature extraction: local binary patterns (LBP) [2], Gabor [19] and scale invariant feature transform (SIFT) [20] features were tested. The second is the convolutional neural network (CNN) based method, where the VGG-Face model [25] was directly adopted. In these experiments, the feature extraction process was the same for caricatures and photos. However, researchers are advocated to design more sophisticated feature extraction methods to solve both the modality variation and misalignment problems.

Matching Algorithm. The last step is to compute the similarity of two sets of features from a caricature and a photo to decide whether they are from the same person. If the misalignment problem is not addressed and the image style variation is not removed, then the matching algorithm should consider removing both alignment and modality variations first. In this case, cross-modality metric learning methods can be useful [11, 12, 13]. Otherwise, if the two problems are fairly well addressed by the previous steps, traditional matching algorithm can be directly applied.

For this step, nine subspace and metric learning methods were tested, including both single modality and cross-modality based methods. For single modality based methods,

principal component analysis (PCA) [15], kernel discriminant analysis (KDA) [4], keep it simple and straightforward metric (KISSME) [17], the information-theoretic metric learning (ITML) [7], the large margin nearest neighbor (LMNN) [30] were tested. For cross-modality based methods, tested methods include canonical correlation analysis (CCA) [9], multi-view discriminant analysis (MvDA) [14], coupled spectral regression (CSR) [18], kernel coupled spectral regression (KCSR) [18]. In the experiments, we demonstrate that removing modality variation by cross-modality based methods, such as CSR and KCSR, can largely improve the performance.

5 Baseline Performance

Following the proposed evaluation protocols and the caricature recognition framework, various alignment methods, feature extraction methods, subspace and metric learning methods were tested on the developed dataset to demonstrate the challenges of caricature recognition and also to provide baseline performances on the WebCaricature dataset. For the following experiments, under the verification settings, adopted evaluation performance measures include VR@FAR=0.1% (denoted as FAR=0.1% in Tables 2, 3, 5), VR@FAR=1% (denoted as FAR=1% in Tables 2, 3, 5) and AUC. Under the identification settings, the evaluation performance measures used include Rank-1 and Rank-10.

5.1 Influence of Alignment Methods

To study the influence of alignment, three alignment methods were tested. The alignment process of the first method was to rotate the image first to align the two eyes to horizontal position. Then, the image was resized to make eye distance of 75 pixels. After this step, the cropped region was defined by making the eye center to the upside of the bounding box by 70 pixels and to the left/right side of the bounding box by 80 pixels. The bounding box is of size 160×200. This alignment method is denoted as eye location based. In the second alignment method, the first two steps were the same. Then, small regions of 40×40 centered at each of the 17 labeled landmarks were also cropped out. This alignment method is denoted as landmark based. The third alignment method was done according to the face contour (defined by the landmarks 1-4). Then the bounding box was enlarged by a scale of 1.2 in both width and height. With the enlarged bounding box, the face image inside the bounding box was resized to 160×200. This alignment method is denoted as bounding box based. The three alignment methods were combined with two feature extraction methods (SIFT and VGG-Face). The resulting feature vectors were combined with PCA for evaluation and all the principal components were kept.

The first 5 rows of Table 2 summarise the results. It can be seen: (1) for SIFT features, it is obvious that landmark based feature extraction (denoted with a suffix '-Land') obtains the best performance. The results of eye location based (denoted by suffix '-Eye') and bounding box based (denoted by suffix '-Box') methods are worse. *In fact, landmark based method is a hard alignment method that forces two patches of images at the same landmark to be directly compared so as to alleviate misalignment problem. Hence it is consistently the best for extracting hand-crafted features, such as SIFT.* (2) For deep features, the bounding box based alignment is much better than the eye location based method. This is perhaps because that the bounding box based alignment can retain more face information. The cropped face images of eye location based method may miss certain part of the face such as chin and

Table 2: Results of various alignment (1-5 rows) and feature extraction methods (6-9 rows) under restricted and unrestricted settings.

Method	Restricted			Unrestricted		
	FAR=0.1% (%)	FAR=1% (%)	AUC	FAR=0.1% (%)	FAR=1% (%)	AUC
SIFT-Eye	2.78 ± 0.46	10.46 ± 1.43	0.738 ± 0.016	2.74 ± 0.75	10.74 ± 1.53	0.734 ± 0.016
SIFT-Land	**4.67 ± 1.08**	**15.39 ± 1.85**	**0.777 ± 0.017**	**4.43 ± 0.82**	**15.24 ± 2.03**	**0.780 ± 0.017**
SIFT-Box	2.96 ± 0.72	9.15 ± 1.17	0.720 ± 0.018	3.42 ± 0.52	11.76 ± 1.21	0.724 ± 0.013
VGG-Eye	21.42 ± 2.02	40.28 ± 2.91	0.896 ± 0.013	19.24 ± 1.95	40.88 ± 2.23	0.898 ± 0.007
VGG-Box	**28.42 ± 2.04**	**55.53 ± 2.76**	**0.946 ± 0.009**	**32.07 ± 2.60**	**56.76 ± 2.35**	**0.946 ± 0.005**
LBP-Eye	0.33 ± 0.15	1.92 ± 0.38	0.600 ± 0.015	0.19 ± 0.04	1.65 ± 0.27	0.597 ± 0.006
Gabor-Eye	3.23 ± 0.74	9.75 ± 1.36	0.716 ± 0.017	2.76 ± 0.47	10.36 ± 1.54	0.718 ± 0.016
SIFT-Eye	2.78 ± 0.46	10.46 ± 1.43	0.738 ± 0.016	2.74 ± 0.75	10.74 ± 1.53	0.734 ± 0.016
VGG-Eye	**21.42 ± 2.02**	**40.28 ± 2.91**	**0.896 ± 0.013**	**19.24 ± 1.95**	**40.88 ± 2.23**	**0.898 ± 0.007**

Table 3: Results of various learning methods under restricted and unrestricted verification settings.

Method	Restricted			Unrestricted		
	FAR=0.1% (%)	FAR=1% (%)	AUC	FAR=0.1% (%)	FAR=1% (%)	AUC
Euc	2.12 ± 0.68	8.28 ± 1.13	0.661 ± 0.014	2.56 ± 0.38	8.40 ± 0.87	0.661 ± 0.009
PCA	4.67 ± 1.08	15.39 ± 1.85	0.777 ± 0.017	4.43 ± 0.82	15.24 ± 2.03	0.780 ± 0.017
KDA	-	-	-	6.62 ± 1.37	24.23 ± 3.26	0.875 ± 0.014
KissME	4.55 ± 1.07	12.15 ± 1.73	0.724 ± 0.011	4.56 ± 0.89	14.66 ± 1.70	0.781 ± 0.016
ITML	**5.08 ± 1.82**	**18.07 ± 4.72**	**0.841 ± 0.018**	5.35 ± 1.20	18.48 ± 2.40	0.828 ± 0.016
LMNN	-	-	-	6.59 ± 1.61	21.37 ± 3.22	0.842 ± 0.014
CCA	4.77 ± 0.68	12.96 ± 1.40	0.775 ± 0.016	5.02 ± 1.19	17.66 ± 2.49	0.812 ± 0.017
MvDA	-	-	-	1.41 ± 0.37	8.29 ± 0.72	0.753 ± 0.014
CSR	-	-	-	**11.76 ± 2.72**	31.86 ± 3.85	0.887 ± 0.013
KCSR	-	-	-	11.66 ± 2.69	**32.00 ± 3.94**	**0.888 ± 0.013**

mouth due to the exaggerations in caricatures. *The deep learning based methods may have mechanisms to alleviate misalignments. Thus the results of bounding box based alignment are better. Note that landmark based method is not applicable for deep learning, but it is straightforward to think that introducing a hard alignment scheme into deep learning would address the misalignment problem.*

5.2   Evaluation of Feature Extraction Schemes

To study the influence of features, LBP, Gabor, SIFT and CNN (VGG-Face) is adopted for experiments, where the first three are hand-crafted features. For all these feature extraction methods, eye location based alignment is used. All the extracted features were combined with PCA for evaluation. In Table 2, 6-9 rows summarizes results. From the table, SIFT feature is almost the best among the hand-crafted features. Gabor is the next. The results of LBP are the worst. The results of VGG-Face are the best, significantly improving over that of SIFT. This illustrates the superiority of deep learned features. Note that VGG-Face was not fine-tuned in this application. Still, the performance is much better than that of any hand-crafted features. *However, under the eye location based alignment setting, the results of VGG-Face for VR@FAR=0.1% and VR@FAR=1% are 21.42% and 40.28% for the restricted setting and 19.24% and 40.88% for the unrestricted setting. Thus, there is still much room for improvement even with deep learning.* Note that there are many other modern deep learning based algorithms that can also be used as baseline. Besides, fine-tuning VGG-Face model is also an interesting approach. These results will be updated in our future work.

Table 4: Results of various learning methods under C2P and P2C identification settings.

Method	C2P		P2C	
	Rank-1 (%)	Rank-10 (%)	Rank-1 (%)	Rank-10 (%)
Euc	13.38 ± 1.10	38.40 ± 1.92	9.04 ± 0.80	29.63 ± 1.35
PCA	15.63 ± 0.82	43.48 ± 1.69	12.47 ± 1.14	40.13 ± 1.53
KDA	19.32 ± 1.36	56.77 ± 1.58	18.92 ± 1.35	57.19 ± 2.61
KissME	15.16 ± 1.63	43.95 ± 2.13	13.30 ± 1.18	43.63 ± 1.64
ITML	15.25 ± 3.07	46.39 ± 6.46	16.48 ± 1.77	49.88 ± 2.29
LMNN	17.92 ± 0.86	50.58 ± 1.72	15.90 ± 1.73	48.08 ± 1.95
CCA	10.84 ± 0.78	40.76 ± 1.08	10.73 ± 0.94	41.12 ± 1.87
MvDA	4.77 ± 0.74	27.73 ± 1.90	4.71 ± 0.87	27.19 ± 2.55
CSR	$\mathbf{25.18 \pm 1.39}$	60.95 ± 1.20	23.36 ± 1.47	60.27 ± 1.97
KCSR	24.87 ± 1.50	$\mathbf{61.57 \pm 1.37}$	$\mathbf{23.42 \pm 1.57}$	$\mathbf{60.95 \pm 2.34}$

5.3 Evaluation of Different Matching Methods

In this section, we tested several single modality subspace learning methods as mentioned in Section 4, including PCA [15], KDA [4] and multi-modality subspace learning methods such as, CCA [9], MvDA [14], CSR [18] and KCSR [18]. The state-of-the-art single modality metric learning methods include the KISSME [17], ITML [7] and LMNN [30]. Landmark based alignment and SIFT features were used for comparing all these methods. Prior to applying these methods, PCA was applied.

As the restricted setting only provides information that two images are either of the same class or not and algorithms such as KDA, LMNN, MvDA, CSR, KCSR require explicit label information for each image, they are not applicable for the image restricted setting. From Table 3, the best result was achieved by ITML under the restricted setting. For the unrestricted setting, the best and second best results were achieved by CSR and KCSR, respectively. *In summary, all these learning methods were better than simple Euclidean distance on original features.* For the unrestricted setting, the performance of CSR and KCSR was the best, because they were designed for cross-modality subspace learning. *This suggests that further studies on cross-modality metric learning or cross-modality subspace learning will be beneficial, due to limited work on this direction.* Under the identification settings, from the results in Table 4, CSR and KCSR achieved similarly the best results. The best rank-1 performance for C2P setting was only 25.18 ± 1.39 and 23.42 ± 1.57 for P2C setting. *This means that there is a large room for improvement on these two settings with the traditional face recognition process.*

5.4 Summary of Results

A summary of the best combinations of the methods at three stages are provided as baselines. Results under verification settings are given in Table 5. For identification settings, the results are summarized in Table 6. From the tables, the deep learning feature based methods outperform the hand-crafted feature based methods to a large extent. Another observation is that the results of VGG-Face can be further improved with KCSR. *This is mainly because that VGG-Face is trained using only photos and may not be able to deal with modality variations. With the help of KCSR to further remove modality variations, the performance can be enhanced. Thus one future direction is to develop end-to-end modality invariant deep learning methods.* Another finding is that performances, *VR@FAR=0.1%, VR@FAR=1%, Rank-1 and Rank-10, of deep learning based methods are still far from satisfactory, indicating that there is still room for improvement.* Lastly, although bounding box based alignment is better

Table 5: Summary of results under restricted and unrestricted settings.

Restricted			
Method	FAR=0.1% (%)	FAR=1% (%)	AUC
SIFT-Land-ITML	5.08 ± 1.82	18.07 ± 4.72	0.841 ± 0.018
VGG-Eye-PCA	21.42 ± 2.02	40.28 ± 2.91	0.896 ± 0.013
VGG-Eye-ITML	18.97 ± 3.90	41.72 ± 5.83	0.911 ± 0.014
VGG-Box-PCA	28.42 ± 2.04	55.53 ± 2.76	0.946 ± 0.009
VGG-Box-ITML	**34.94 ± 5.06**	**57.22 ± 6.50**	**0.954 ± 0.010**
Unrestricted			
Method	FAR=0.1% (%)	FAR=1% (%)	AUC
SIFT-Land-KCSR	11.66 ± 2.69	32.00 ± 3.94	0.888 ± 0.013
VGG-Eye-PCA	19.24 ± 1.95	40.88 ± 2.23	0.898 ± 0.007
VGG-Eye-KCSR	23.46 ± 2.65	48.57 ± 2.73	0.925 ± 0.007
VGG-Box-PCA	32.07 ± 2.60	56.76 ± 2.35	0.946 ± 0.005
VGG-Box-KCSR	**39.09 ± 3.62**	**65.82 ± 2.48**	**0.963 ± 0.004**

Table 6: Summary of results under C2P and P2C settings.

C2P		
Method	Rank-1 (%)	Rank-10 (%)
SIFT-Land-KCSR	24.87 ± 1.50	61.57 ± 1.37
VGG-Eye-PCA	35.07 ± 1.84	71.64 ± 1.32
VGG-Eye-KCSR	39.76 ± 1.60	75.38 ± 1.34
VGG-Box-PCA	49.89 ± 1.97	84.21 ± 1.08
VGG-Box-KCSR	**55.41 ± 1.41**	**87.00 ± 0.92**
P2C		
Method	Rank-1 (%)	Rank-10 (%)
SIFT-Land-KCSR	23.42 ± 1.57	60.95 ± 2.34
VGG-Eye-PCA	36.18 ± 3.24	68.95 ± 3.25
VGG-Eye-KCSR	40.67 ± 3.61	75.77 ± 2.63
VGG-Box-PCA	50.59 ± 2.37	82.15 ± 1.31
VGG-Box-KCSR	**55.53 ± 2.17**	**86.86 ± 1.42**

than eye based alignment, as analyzed in Section 5.1, *there is still lack of good alignment methods of caricatures for deep learning.*

6   Conclusions

A new benchmark dataset of face caricatures and photos is presented, together with a framework and protocols, to facilitate caricature recognition and tackling its challenges. The main contribution of this paper includes: a large caricature dataset of 252 people with 6024 caricatures and 5974 photos, which is made publicly available. Facial landmarks, evaluation protocols and baseline performances are provided on the dataset. Following these protocols and the framework, a set of face alignment methods, hand-crafted and deep learning features, and various subspace and metric learning methods are tested. A conclusion is that there is still room for improvement even with the best results. With this dataset and from the baseline evaluations, there are several future directions. Caricature face landmark detection is of great interest and a key step for caricature recognition. As the performance on this dataset is still far from saturated, future work on caricature and face feature extraction and cross-modal metric learning methods are also promising directions.

Acknowlegement

References

[1] Bahri Abaci and Tayfun Akgul. Matching caricatures to photographs. *Signal, Image and Video Processing*, 9(1):295–303, 2015.

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns:

Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2007.

[3] Francis R Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9(Jun):1179–1225, 2008.

[4] Deng Cai, Xiaofei He, and Jiawei Han. Speed up kernel discriminant analysis. *VLDB J.*, 20(1):21–33, 2011.

[5] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[6] Elliot J. Crowley, Omkar M. Parkhi, and Andrew Zisserman. Face painting: querying art with photos. In *Proc. Brit. Mach. Vis. Conf.*, pages 65.1–65.13, September 2015.

[7] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proc. Int. Conf. Mach Learn.*, pages 209–216, 2007.

[8] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[9] Harold Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936.

[10] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[11] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin. Heterogeneous face recognition by margin-based cross-modality metric learning. *IEEE Trans. Cybern.*, 48(6):1814–1826, 2018.

[12] J. Huo, Y. Gao, Y. Shi, and H. Yin. Cross-modal metric learning for auc optimization. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–13, 2018. ISSN 2162-237X. doi: 10.1109/TNNLS.2017.2769128.

[13] Jing Huo, Yang Gao, Yinghuan Shi, and Hujun Yin. Variation robust cross-modal metric learning for caricature recognition. In *Proc. Thematic Workshops ACM Multimedia*, pages 340–348. ACM, 2017.

[14] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):188–194, 2016.

[15] Kyungnam Kim. Face recognition using principle component analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 586–591, 1996.

[16] Brendan F Klare, Serhat S Bucak, Anil K Jain, and Tayfun Akgul. Towards automated caricature recognition. In *Proc. Int. Conf. Biometrics*, pages 139–146, 2012.

[17] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2288–2295, 2012.

[18] Zhen Lei and Stan Z Li. Coupled spectral regression for matching heterogeneous faces. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1123–1128, 2009.

[19] Chengjun Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Process.*, 11 (4):467–476, 2002.

[20] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, Nov 2004.

[21] Robert Mauro and Michael Kubovy. Caricature and face recognition. *Memory & Cognition*, 20(4):433–440, 1992.

[22] MegviiInc. Face++ research toolkit. www.faceplusplus.com, December 2013.

[23] Nandan Rai S. Mishra A. Mishra, A. and C. V. Jawahar. Iiit-cfw: A benchmark database of cartoon faces in the wild. In *VASE ECCVW*, pages 35–47, 2016.

[24] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, and Xueming Li. Cross-modal face matching: beyond viewed sketches. In *Proc. Asian Conf. Comput. Vis.*, pages 210–225, 2014.

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. Brit. Mach. Vis. Conf.*, 2015.

[26] David Perkins. A definition of caricature and caricature and recognition. *Studies in Visual Communication 2.1*, pages 1–24, 1975.

[27] Jobany Rodriguez and Ricardo Gutierrez-Osuna. Reverse caricatures effects on three-dimensional facial reconstructions. *Image Vis. Comput.*, 29(5):329–334, 2011.

[28] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proc. IEEE*, 94(11):1948–1962, 2006.

[29] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1988–1996, 2014.

[30] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10(Feb):207–244, 2009.