Towards Automated Caricature Recognition

Brendan F. Klare, Serhat S. Bucak, and Anil K. Jain*
Michigan State University
East Lansing, MI, U.S.A.

Tayfun Akgul
Istanbul Technical University
Maslak, Istanbul, Turkey

Abstract

This paper addresses the problem of identifying a subject from a caricature. A caricature is a facial sketch of a subject's face that exaggerates identifiable facial features beyond realism, while still conveying his identity. To enable this task, we propose a set of qualitative facial features that encodes the appearance of both caricatures and photographs. We utilized crowdsourcing, through Amazon's Mechanical Turk service, to assist in the labeling of the qualitative features. Using these features, we combine logistic regression, multiple kernel learning, and support vector machines to generate a similarity score between a caricature and a facial photograph. Experiments are conducted on a dataset of 196 pairs of caricatures and photographs, which we have made publicly available. Through the development of novel feature representations and matching algorithms, this research seeks to help leverage the ability of humans to recognize caricatures to improve automatic face recognition methods.

1. Introduction

Among the remarkable capabilities possessed by the human visual system, perhaps none is more compelling than our ability to recognize a person from a caricature. A caricature is a face image in which certain facial attributes and features have been exaggerated to a degree that is often beyond realism, and yet the face is still recognizable (see Fig. 1). As Leopold et al. discussed [20], the caricature generation process can be conceptualized by considering each face to lie in a face space. In this space, a caricature face beyond the line connecting the mean face[1] and a subject's face. In other words, a caricature is an extrapolated version of the original face.

Despite the (often extreme) exaggeration of facial features, the identity of a subject in a caricature is generally obvious, provided the original face is known to the viewer. In fact, studies have suggested that people may be better at recognizing a familiar person through a caricature portrait than from a veridical portrait[2] [23, 26].

So why is it that an exaggerated, or extrapolated, version of a face can be so easy to recognize? Studies in human cognition have suggested this phenomenon is correlated to how humans represent and encode facial identity [23]. Empirical studies suggest that this representation involves the use of prototype faces, where a face image is encoded in terms of its similarity to a set of prototype face images [28, 31, 20]. Under this assumption, the effectiveness of a caricature would be due to its ability to emphasize deviations from prototypical faces. This would also explain why faces that are "average" looking, or typical, are more difficult to recognize [31].

Automated face recognition, despite its significant progress over the past decade [10], still has many limitations. State of the art face recognition algorithms are not able to meet the performance requirements in uncontrolled and non-cooperative face matching scenarios, such as surveillance. We believe clues on how we can better compute the similarity between faces may be found through investigating the caricature matching process.

In this paper, we study the process of automatically matching a caricature to a facial photograph. To accomplish this, we define a set of qualitative facial attributes (e.g. "nose to mouth distance") that are used to encode the appearance of a face (caricature or photograph). These features, called "qualitative features", are generally on a nominal scale (and occasionally on an ordinal scale) that characterize when a particular facial attribute is either typical or atypical (deviates from the mean face). Statistical learning is performed to learn feature weighting and the optimal subset of these features.

While several methods exist for automating the caricature generation process, to our knowledge, this is the first attempt to automate the caricature recognition process. In addition to posting impressive accuracies on this difficult heterogeneous face recognition task, we are also releasing a caricature recognition dataset, experimental protocol, and

*A.K. Jain is also with the Dept. of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea.

[1] A mean face is the average appearance of all faces.

[2] A verdical portrait is a highly accurate facial sketch of a subject.

978-1-4673-0397-2/12/$31.00 ©2012 IEEE

(a) (b) (c) (d)

Figure 1. Examples of caricatures (top row) and photographs (bottom row) of four different personalities. Shown above are: (a) Angelina Jolie (drawn by Rok Dovecar), (b) Adam Sandler (drawn by Dan Johnson), (c) Bruce Willis (drawn by Jon Moss), and (d) Taylor Swift (drawn by Pat McMichael).

qualitative features to the research community. Through the design and performance evaluation of caricature recognition algorithms, it is our belief that we will help advance the state of automatic face recognition through the discovery of additional facial representations and feature weightings [2].

2. Related Work

Caricature recognition belongs to a face recognition paradigm known as heterogeneous face recognition (HFR) [14]. Heterogeneous face recognition is the task of matching two faces from alternate modalities.

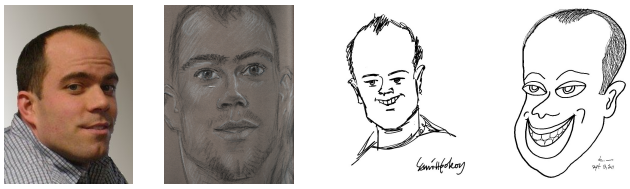
Solutions to heterogeneous face recognition problems generally follow one of two approaches. The first approach, popularized by Wang and Tang [32], seeks to synthesize an image from one of the modalities (e.g. sketch) in the second modality (e.g. photograph). Once this synthesis has occurred, standard matching algorithms can be applied in the now common modality.

The second approach to HFR is to densely sample feature descriptors (such as local binary patterns (LBP) [24]) from the images in each modality. The feature descriptor is selected such that it varies little when moving between the imaging modalities, while still capturing key discriminative information. A benefit of this feature-based approach is that it facilitates statistical subspace learning (such as linear discriminant analysis (LDA) [6] and its variants) to further improve the class separability. This approach has been successfully used by Liao et al. [22], Klare and Jain [16, 14, 13], and Bhatt et al. [7].

In the context of caricature recognition, an image feature descriptor-based approach is challenged because the descriptors from the caricature and photograph may not be highly correlated due to misalignment caused by feature exaggerations (e.g. the nose in the caricature may extend to where the mouth or chin is in the photograph). However, the application of LDA, in a manner similar to other HFR studies [16, 13], somewhat compensates for these misalignments. Further, LDA offers a solution to the intra-artist variability through the modeling of the within-class scatter. For these reasons, this paper also makes use of the image feature descriptor-based approach in addition to the qualitative feature based approach (see Section 6).

A major contribution of this paper is to define a set of categorical, or nominal, facial attributes. This approach is similar to the attribute and simile features proposed by Kumar et al. [18], who demonstrated the benefit of this nominal feature representation for recognizing face images. While we present a similar representation, the features proposed here have been carefully defined by a professional artist with experience in drawing caricatures.

A number of methods in graphics have been developed for automatically generating caricature images [8, 3, 4, 17, 21]. However, to our knowledge, no previous research on matching caricatures to photographs has been conducted. The method proposed by Hsu and Jain [11] was the closest attempt, where facial photographs were matched by first



(a) (b) (c) (d)

Figure 2. Different forms of facial sketches (b-d). (a) Photograph of subject. (b) Portrait sketch. (c) Forensic sketch drawn by Semih Poroy from a verbal description. (d) Caricature sketch.

synthesizing them into caricature drawings.

3. Caricature Dataset

In this section we describe the dataset that was used in this study[3]. Future studies comparing accuracies on this dataset should follow the protocol detailed in Section 7.

The dataset consists of pairs of a caricature sketch and a corresponding facial photograph from 196 subjects (see Fig. 1 for examples). Two sources were used to collect these images. The first was through contacts with various artists who drew the caricatures. For these images, permission was granted to freely distribute the caricatures. In total 89 caricatures were collected from this source.

The second source of caricature images was from Google Image searches. For these caricatures, the url of the image was recorded, and is included in the dataset release (along with the actual image). There were 107 pairs from this source.

The corresponding face image for each subject was provided by the caricature artist for caricatures from the first source, and by Google Image search for the second source. When selecting face photographs, care was taken to find images that had minimal variations in pose, illumination, and expression. However, such "ideal" images do not always exist. Thus, many of the PIE (pose, illumination and expression) factors still persist.

4. Qualitative Feature Representation

We define a set of categorical facial features for representing caricature images and face photographs. These features were developed by one of the authors who is a cartoonist (in addition to being a professor of electrical engineering [1]).

While face images are typically encoded by high-dimensional numerical features (such as local binary patterns [24]), the tendency of a caricature to exaggerate distinctive facial features [25] makes such numerical encodings not appropriate for representing caricatures images. Instead, the proposed qualitative features describe facial features that a caricature artist may portray as to whether or not it is present. Thus, if "large distance between the nose and mouth" is a feature the artist chooses to emphasize, the proposed representation is able to capture this without being impacted by exactly how much the artist extrapolates this distance from the norm [2].

A caricaturist can be likened to a "filter" that only retains useful information in a face for identification. As a filter, the artist uses his talent to analyze a face, eliminate insignificant facial features, and capture the identity though exaggeration of the prominent features. Most of the caricaturists start with the description of the general shape of the head. They assemble the eyes, nose, eyebrows, lips, chin and ears with some exaggerations in geometrically correct locations (always maintaining the appropriate ratios amongst them); finally, they include the hair, moustache and beard (depending on the gender or their presence in the face).

In this study, following the caricaturists methodology [25], we define a set of 25 qualitative facial features that are classified into two levels (see Figure 3). The first level (Level 1) is defined for the general shapes and sizes of the facial components and the second level (Level 2) is defined for the size and appearance of facial components, as well as ratios amongst the locations of different components (e.g. distance of the mouth from the nose).

4.1. Level 1 Qualitative Features

Level 1 features describe the general appearance of the face. These features can be more quickly discerned than Level 2 features. In standard face recognition tasks, Level 1 features are less informative than Level 2 features [15] due to their lack of persistence and uniqueness. However, in caricature recognition experiments these features are shown to be the most informative (see Section 7).

The length of the face is captured by the Face Length feature (narrow or elongated). The shape of the face is described by the Face Shape feature (boxy, round, or triangular). Two different features are used to capture the hair style, with values including: short bangs, parted left, parted right, parted middle, bald, nearly bold, thin middle, and curly. Facial hair is represented with the Beard feature (none, normal, Abraham Lincoln, thin, thick, and goatee) and Mustache feature (normal, none, thin, and thick) features. See Figure 3 for visual examples of these features.

4.2. Level 2 Features

Specific facial details are captured by the Level 2 facial features. Level 2 facial features will offer more precise descriptions of specific facial components (such as the eyes, nose, etc.) compared to Level 1 features.

Several binary features are used to represent the appearance of the eyes. These include whether or not the eyes are dark, sleepy, "almond" shaped, slanted, sharp, or baggy. Similarly, the eyebrows are represented by their thickness,

[3]Dataset available by sending a request to tayfunakgul@itu.edu.tr

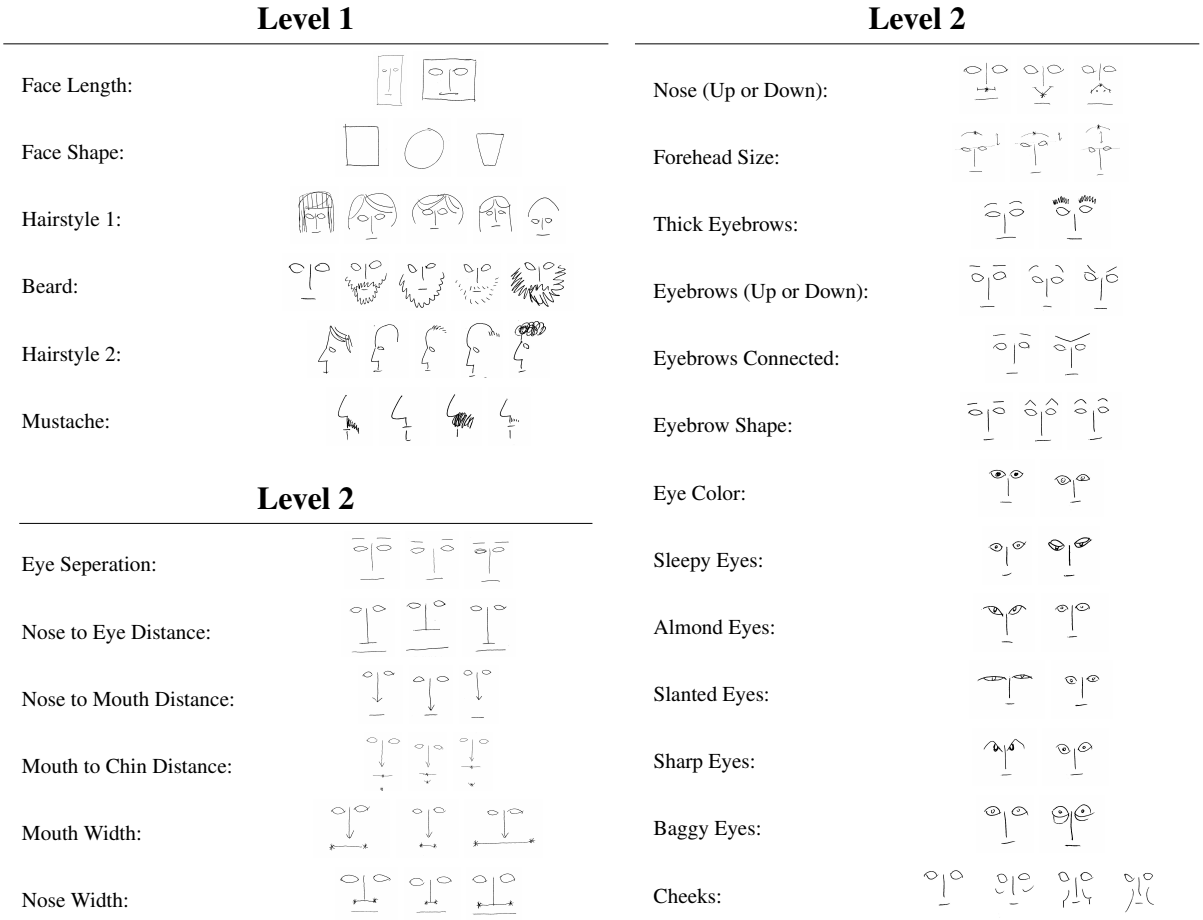


Figure 3. Illustration of the twenty five qualitative features used to represent both caricatures and photographs. The similarity between sketches and photographs were measured within this representation.

connectedness, direction (up or down), and general shape (normal, rounded, or pointed). The nose is represented by its width (normal, thin or wide) and direction (normal, points up, or points down). The mouth is characterized by its width (normal, thin, or wide). The cheeks are described as being either normal, thin, fat or baggy.

Several features are used to capture the geometric relationships among the facial components. They describe the distance between the nose and the eyes, the nose and the mouth, and the mouth and the chin. Two additional features describe the distance between the eyes, and the length of the forehead.

4.3. Feature Labeling

Each image (caricature and photo) was labeled with qualitative features by annotators provided through Amazon's Mechanical Turk[4]. Several annotators combined to label the entire set of image pairs with each of the 25 qualitative features. Each annotator was asked to label a single image with a single feature value at a time. Thus, the annotator was shown an image of either a caricature or a photograph, and each of the possible feature values (along with their verbal description) for the current feature being labeled.

To compensate for differences in annotator opinions on less obvious image/feature combinations, each image was labeled three times by three different annotators. Thus, given 25 qualitative features and three labelers per feature, a total of 75 feature labels were available per image. In all, 29,400 labeling tasks were performed through this crowdsourcing method (costing roughly $300 USD).

5. Matching Qualitative Features

With each image labeled with 25 qualitative attributes u times ($u = 3$, see Sec. 4.3), each image (photo or caricature) can be represented by a $u \times 25$ matrix $\mathbf{C} \in Z_+^{u \times 25}$. Note that the matrix elements are nonnegative integers since each feature is of categorical type.

In order to improve the matching performance, we adopt

[4] https://www.mturk.com/

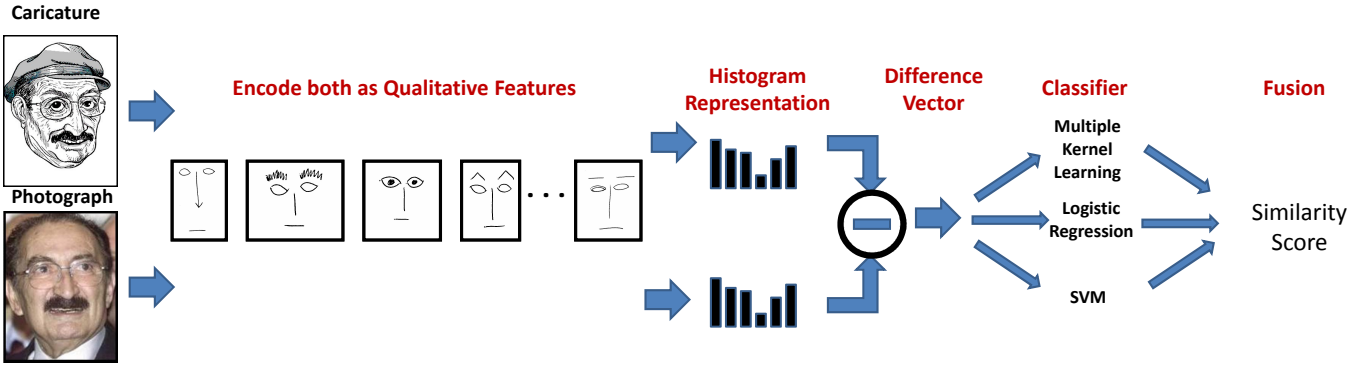


Figure 4. Overview of the caricature recognition algorithm.

machine learning techniques for feature subset selection and weighting. To facilitate this, we encode the categorical attributes into binary features by using r_i bits for each attribute, where r_i is the number of possible choices for the i^{th} attribute. For example, $r_i = 2$ for "*Thick Eyebrows*" and $r_i = 3$ for "*Nose Width*" (see Figure 3).

Ideally, the binary valued feature vector should lead to a vector with only one non-zero element per feature. However, the annotators may give contradicting annotations (e.g. one subject can be labeled as having a "Wide Nose" and "Normal Nose" by two different annotators). Hence, we accumulate the binary valued feature vectors into histogram feature vectors. Thus, a single feature will no longer be represented by an r_i-bit binary number, but instead by an r_i-dimensional feature vector. Each component will have a minimum value of 0 and a maximum value of u. Finally, for each image, we concatenate the 25 individual attribute histograms to get a 77-dimensional feature vector ($\mathbf{x} \in \mathbb{Z}_+^{77}, ||\mathbf{x}||_1 = 25u$). Given this representation, the simplest method for matching is to perform nearest neighbor search with Euclidean distance (referred to as NN_{L_2}).

Next, we convert the caricature-photo matching problem into a binary classification task by calculating the absolute difference vector for every possible caricature-photo pair in the training set. In the binary classification setting, the difference vector for the caricature and photo pair of the same subject (i.e. a true match) is labeled as '1' whereas the difference vector for caricature-photo pair of two different subjects (i.e. a false match) is labeled as '-1'. This gives us n positive samples (genuine matches) and $n^2 - n$ negative samples (imposter matches), where n is the number of subjects in the training set.

With the caricature recognition problem reformulated as a binary classification task, we leverage a fusion of several binary classifiers. Let $\{(\mathbf{x_i}, y_i), \mathbf{x_i} \in \mathbb{R}^d, y_i \in \{-1, 1\}, i = 1, 2, ..., m\}$ be the m pairs of difference vectors, where $d = 77$. Again, if $\mathbf{x}_i$ is a difference vector between a caricature and photograph of the same subject then $y_i = 1$, otherwise $y_i = -1$.

5.1. Logistic Regression

Logistic regression seeks to find a function that maps the difference vectors to their numerical label ($+1$ or -1). The output of this regression can be interpreted as a similarity score, which facilitates fusion and receiver operator characteristic (ROC) analysis.

The objective function of the logistic regression is as follows

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{m} \{-y_i \mathbf{x_i}'\boldsymbol{\beta} + \log(1 + \exp(\mathbf{x_i}'\boldsymbol{\beta}))\} + \lambda R(\boldsymbol{\beta}), \quad (1)$$

where $\boldsymbol{\beta}$ is the vector of the feature weights to be learned, $R(\boldsymbol{\beta})$ is a regularization term (to avoid overfitting and impose structural constraints) and λ is a coefficient to control the contribution of the regularizer to the cost. Two different regularizers are commonly used: (i) the L_1-norm regularizer, $R(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1$ (also know as Lasso [30]), which imposes sparseness on the solutions by making most of the coefficients to be equal to zero for large values of λ, and (ii) the L_2-norm regularizer, $R(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2$, which leads to non-sparse solutions.

Our experimental results with the implementation of [12] favored the L_2-norm regularizer, which we refer to in Section 7 as *Logistic Regression*. Having solved for $\boldsymbol{\beta}$ using a gradient descent method, we compute the similarity value of the difference vector $\mathbf{x}$ between a caricature and photograph as: $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}\boldsymbol{\beta}))$.

5.2. Multiple Kernel Learning and SVM

One limitation of the logistic regression method is that it is restricted to finding linear dependencies between the features. In order to learn non-linear dependencies we use support vector machines (SVM) and multiple kernel learning (MKL) [5].

Given m training images, we let $\{\mathbf{K}_j \in \mathbb{R}^{m \times m}, j = 1, \ldots, 25\}$ represent the set of base kernels. $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_s)^\top \in \mathbb{R}_+^s$ denotes the coefficients used to combine these base kernels, and $\mathbf{K}(\mathbf{p}) = \sum_{j=1}^{s} \mathbf{p}_j \mathbf{K}_j$ is the

Method	TAR @ FAR=10.0%	TAR @ FAR=1.0%	Rank-1	Rank-10
Qualitative Features (no learning):				
NN_{L_2}	39.2 ± 5.4	9.4 ± 2.7	12.1 ± 5.2	52.1 ± 7.1
Qualitative Features (learning):				
Logistic Regression	50.3 ± 2.4	11.3 ± 2.9	17.7 ± 4.2	62.1 ± 3.8
MKL	39.5 ± 3.2	7.4 ± 3.9	11.0 ± 3.9	50.5 ± 4.0
NN_{MKL}	46.6 ± 3.9	10.3 ± 3.6	14.4 ± 2.9	59.5 ± 3.9
SVM	52.6 ± 5.0	12.1 ± 2.8	20.8 ± 5.6	65.0 ± 3.8
Logistic Regression+NN_{MKL}+SVM	56.9 ± 3.0	15.5 ± 4.6	23.7 ± 3.5	70.5 ± 4.4
Image Descriptors (learning):				
LBP with LDA	33.4 ± 3.9	11.5 ± 2.5	15.5 ± 4.6	42.6 ± 4.6
Qualitative Features + Image Descriptors:				
Logistic Regression+NN_{MKL}+ SVM+LBP with LDA	61.9 ± 4.5	22.7 ± 3.5	32.3 ± 5.1	74.8 ± 3.4

Table 1. Average identification and verification accuracies of the proposed qualitative, image feature-based, and baseline methods. Average accuracies and standard deviations were measured over 10 random splits of 134 training subjects and 62 testing subjects (subjects in training and test sets are different).

combined kernel matrix. We learn the coefficient vector $\mathbf{p}$ by solving the convex-concave optimization of the MKL dual formulation [19]:

$$\min_{\mathbf{p}\in\Delta}\max_{\alpha\in\mathcal{Q}} \widehat{\mathbf{L}}(\boldsymbol{\alpha},\mathbf{p}) = \mathbf{1}^\top\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})^\top\mathbf{K}(\mathbf{p})(\boldsymbol{\alpha}\circ\mathbf{y}), \quad (2)$$

where $\circ$ denotes the Hadamard (element-wise) product, $\mathbf{1}$ is a vector of all ones, and $\mathcal{Q} = \{\boldsymbol{\alpha} \in [0, C]^m\}$ is the domain for dual variables $\boldsymbol{\alpha}$. Note that this formulation can be considered as the dual formulation of SVM for the combined kernel.

One popular choice for domain Δ is $\Delta_2 = \left\{\mathbf{p} \in \mathbb{R}^s_+ : \|\mathbf{p}\|_2 \leq 1\right\}$. Often the L_1 norm is used to generate a sparse solution, however, in our application, the small sample size impacted the accuracy of this approach.

For MKL, each individual attribute is considered as a separate feature by constructing one kernel for each attribute (resulting in 25 base kernels). Our MKL classifier was trained using an off-the-shelf MKL tool [29].

Once this training is complete, we are able to measure the similarity of a caricature and photograph (represented as the difference vector $\mathbf{x}$) by: $f(\mathbf{x}) = \sum_{i=1}^{n_s} \boldsymbol{\alpha}_i y_i \mathbf{K}_p(\mathbf{x}_i, \mathbf{x})$, where n_s is the number of support vectors, and $\mathbf{K}_p(\cdot)$ is the combined matrix. In Section 7, we refer to this method as *MKL*.

In addition to the MKL algorithm, we also use the standard SVM algorithm [9] by replacing the multiple kernel matrix $\mathbf{K}_p(\cdot)$ with a single kernel $\mathbf{K}(\cdot)$ that utilizes all feature components together. In Section 7, we refer to this approach as *SVM*. Both the MKL and SVM algorithms used RBF kernels (the kernel bandwidth was determined empirically).

Finally, we introduce a method known as the nearest neighbor MKL (NN_{MKL}). Because the vector $\mathbf{p}$ in Eq. 2 assigns weight to each of the 25 qualitative features, we can explicitly use these weights to perform weighted nearest neighbor matching. Thus, the dissimilarity between a caricature and photograph is measured as the sum of weighted differences between each of the qualitative feature vectors.

6. Image Descriptor-based Recognition

As discussed, encoding facial images with low level feature descriptors such as local binary patterns [24] is challenged with respect to matching caricatures to photograph due to the misalignments caused from the feature exaggeration in caricatures. However, since this approach has seen success in matching facial sketches to photographs [16, 7, 14], we also employ a similar technique for the caricature matching task.

The first step in the image descriptor-based algorithm is to align the face images using the two eye locations. These locations are marked manually due to the wide variations of pose in both the photographs and (especially) the caricatures. Using the center of the eyes, we performed planar rotation to fix the face upright, scaled the image to 75 pixels between the eyes, and cropped the image to a height of 250 pixels and a width of 200 pixels.

For both caricatures and photographs, we densely sampled local binary pattern histograms from image patches of 32 by 32 pixels. Next, all of the LBP histograms computed from a single image are concatenated into a single feature

vector. Finally, we performed feature-based random subspace analysis [13] by randomly sampling the feature space b times. For each of the b subspaces, linear discriminant analysis (LDA) is performed to extract discriminative feature subspaces [6]. In Section 7 we will refer to this method as *LBP with LDA*.

7. Experimental Results

The 196 pairs of caricatures and photographs (see Section 3), were randomly split such that 134 pairs (roughly 2/3rd) were made available for training, and 62 pairs (roughly 1/3rd) was available for testing. These sets were non-overlapping (i.e. no subject used in training was used for testing). We partitioned the data into training and testing sets 10 different times, resulting in 10 different matching experiments. The results shown in this section are the mean and standard deviation of the matching accuracies from those 10 random partitions. The precise splits used for these experiments are included with the release of the caricature image dataset.

The performance of each matching algorithm was measured using both the cumulative match characteristic (CMC) and the receiver operating characteristic (ROC) curves. For the CMC scores, we list the Rank-1 and Rank-10 accuracies. With 62 subjects available for testing, the gallery size was 62 images (photographs), and the scores listed are the average rank retrieval when querying the gallery with the 62 corresponding caricatures. The ROC analysis is listed as the true accept rate (TAR) at fixed false accept rates (FAR) of 1.0% and 10.0%.

Table 1 lists the results for each of the recognition algorithms discussed in this work. Even without learning the qualitative features (NN_{L_2}) still had a higher accuracy than the image descriptor-based method (LBP with LDA). Thus, while image descriptor-based methods work well in matching verdical sketches to photographs [16], the misalignments caused by the exaggerations in the caricatures challenge this method. At a false accept rate of 10.0%, several of the proposed learning methods (*Logistic Regression,* NN_{MKL}*, and SVM*) are able to improve the accuracy of the qualitative features by around 10%. Despite the inability of the *MKL* method to improve the matching accuracy, using the weights from MKL with the nearest neighbor matcher (NN_{MKL}) improves the matching accuracy.

Because the classification algorithms used in this study output numerical values that indicate the similarity of a caricature image and a photograph, we are able to leverage fusion techniques to further improve the accuracy. Fusion of algorithms in Table 1 are denoted by the a '+' symbol between algorithms names. This indicates the use of sum of score fusion with min-max score normalization [27].

Using only qualitative features, the matching accuracy (at FAR=10.0%) was improved to nearly 57% (using *Logistic Regression+*NN_{MKL}*+SVM*). While the image descriptor-based method performed poorly with respect to the qualitative features, it proved valuable when added to the fusion process: *Logistic Regression+*NN_{MKL}*+SVM+LBP with LDA* had an accuracy of 61.9%.

Feature Name	Weight	Feature Name	Weight
Hairstyle 1	*2.86*	*Almond Eyes*	*0.21*
Beard	*0.85*	*Nose (Up or Down)*	*0.21*
Mustache	*0.81*	*Face Shape*	*0.20*
Hairstyle 2	*0.70*	*Forehead Size*	*0.19*
Eyebrows (Up or Down)	*0.45*	*Eye Color*	*0.18*
Nose to Mouth Distance	*0.43*	*Sleepy Eyes*	*0.14*
Eye Seperation	*0.43*	*Sharp Eyes*	*0.13*
Nose Width	*0.42*	*Baggy Eyes*	*0.12*
Face Length	*0.27*	*Nose to Eye Distance*	*0.12*
Cheeks	*0.27*	*Thick Eyebrows*	*0.11*
Mouth Width	*0.26*	*Eyebrows Connected*	*0.10*
Mouth to Chin Distance	*0.23*	*Slanted Eyes*	*0.10*
Eyebrow Shape	*0.22*		

Figure 5. The multiple kernel learning (MKL) weights ($\mathbf{p}$), scaled by 10, for each of the qualitative features. Higher weights indicate more informative features.

Using the estimated $\mathbf{p}$ vector in the multiple kernel learning (MKL) algorithm, we are able to interpret the relative importance of each of the qualitative features. Since each component of $\mathbf{p}$ corresponds to the weight assigned to each of the 25 qualitative features, we can loosely interpret this vector to understand which features provided the most discriminative information. Figure 5 lists the weights for each of the 25 facial features. Surprisingly, we see that the Level 1 qualitative features are more discriminative than the Level 2 facial features. While this is counter to a standard face recognition task [15], caricatures are different in nature than face images. We believe the relative importance of Level 1 facial features in this setting is akin to the information an artist filters from the face.

8. Summary

This paper introduces a challenging new problem in heterogeneous face recognition: matching facial caricatures to photographs. We have defined a set of qualitative facial features for representing both caricatures and photographs. A suite of statistical learning algorithms are adopted to learn the most salient combinations of these features from a training set of caricature and photograph pairs.

In order to facilitate research in caricature matching, we are releasing the initial dataset of 196 pairs of caricatures and photographs used in this study. We believe progress towards automating the recognition of caricatures will profoundly impact the progress in face recognition, particularly

in improving facial representations and indexing.

Our future efforts in caricature recognition will be to increase the size of the caricature database, and utilize what we have learned here to augment face recognition performance.

Acknowledgements

We would like to acknowledge the help of many different renowned caricaturists and cartoonists, especially Semih Poroy, Gungor Kabakcioglu (who recently passed away) and Hakan Celik. We would like to thank Dr. Ergun Akleman for his contribution. Anil Jain's research was also partially supported by the World Class University program funded by the Ministry of Education, Science and Technology through the National Research Foundation of Korea (R31-10008).

References

[1] T. Akgul. Introducing the cartoonist, Tayfun Akgul. *IEEE Antennas and Propagation Magazine*, 49(3):162, 2007. 3

[2] T. Akgul. Can an algorithm recognize montage portraits as human faces? *IEEE Signal Processing Magazine*, 28(1):160 –158, 2011. 2, 3

[3] E. Akleman. Making caricatures with morphing. In *Proc. ACM SIGGRAPH*, 1997. 2

[4] E. Akleman. Modeling expressive 3d caricatures. In *Proc. ACM SIGGRAPH*, 2004. 2

[5] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008. 5

[6] P. Belhumeur, J. Hespanda, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 19(7):711–720, 1997. 2, 7

[7] H. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. On matching sketches with digital face images. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, pages 1 –7, 2010. 2, 6

[8] S. Brennan. Caricature generator: The dynamic exaggeration of faces by computer. *Leonardo*, 18:170–178, 1985. 2

[9] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. 6

[10] P. J. Grother, G. W. Quinn, and P. J. Phillips. MBE 2010: Report on the evaluation of 2d still-image face recognition algorithms. *National Institute of Standards and Technology, NISTIR*, 7709, 2010. 1

[11] R.-L. Hsu and A. Jain. Generating discriminating cartoon faces using interacting snakes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(11):1388 – 1398, 2003. 2

[12] Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Staistica Sinica*, 16:375–390, 2006. 5

[13] B. Klare and A. Jain. Heterogeneous face recognition: Matching NIR to visible light images. In *Proc. International Conference on Pattern Recognition*, 2010. 2, 7

[14] B. Klare and A. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Trans. Pattern Analysis & Machine Intelligence (under review)*, 2011. 2, 6

[15] B. Klare and A. K. Jain. On a taxonomy of facial features. In *Proc. IEEE Conference on Biometrics: Theory, Applications and Systems*, 2010. 3, 7

[16] B. Klare, Z. Li, and A. Jain. Matching forensic sketches to mugshot photos. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2011. 2, 6, 7

[17] H. Koshimizu, M. Tominaga, T. Fujiwara, and K. Murakami. On kansei facial image processing for computerized facial caricaturing system picasso. In *Proc. IEEE Conference on Systems, Man, and Cybernetics*, 1999. 2

[18] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(10):1962 –1977, 2011. 2

[19] G. Lanckriet, N. Cristianini, P. Bartlett, and L. E. Ghaoui. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004. 6

[20] D. A. Leopold, A. J. O'Toole, T. Vetter, and V. Blanz. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, 4:89–94, 2001. 1

[21] T. Lewiner, T. Vieira, D. Martinez, A. Peixoto, V. Mello, and L. Velho. Interactive 3d caricature from harmonic exaggeration. *Computers and Graphics*, 35(3):586–595, 2011. 2

[22] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Li. Heterogeneous face recognition from local structures of normalized appearance. In *Proc. Int. Conference on Biometrics*, 2009. 2

[23] R. Mauro and M. Kubovy. Caricature and face recognition. *Memory & Cognition*, 20(4):433–440, 1992. 1

[24] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 24(7):971–987, 2002. 2, 3, 6

[25] L. Redman. *How to draw caricatures*. McGraw-Hill, 1984. 3

[26] G. Rhodes, S. Brennan, and S. Carey. Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, 19(4):473–497, 1987. 1

[27] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, 2003. 7

[28] R. L. Solso and J. E. McCarthy. Prototype formation of faces: A case of pseudo-memory. *British Journal of Psychology*, 72(4):499–503, 1981. 1

[29] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006. 6

[30] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994. 5

[31] T. Valentine and V. Bruce. The effects of distinctiveness in recognising and classifying faces. *Perception*, 15(5):525–535, 1986. 1

[32] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 31(11):1955–1967, Nov. 2009. 2