An Empirical Investigation on the Effect of Shape Exaggeration in Face Sketch to Photo Matching

1Samsul Setumin, 2Abduljalil Radman, 3Shahrel Azmin Suandi
1,2,3Intelligent Biometric Group, School of Electrical and Electronics Engineering, Universiti Sains Malaysia, Engineering Campus, Nibong Tebal 14300, Malaysia.
1Faculty of Electrical Engineering, Universiti Teknologi MARA, Cwangan Pulau Pinang, Permatang Pauh 13500, Malaysia.
Email: 1samsul@student.usm.my, 1samsuls@uitm.edu.my, 2abdural@usm.my, 3shahrel@usm.my

Abstract—Facial sketch to photo retrieval system has useful application in forensic investigation. The main problem identified in the system is that the sketch and photo are from different modality. In order to address this problem, intra-modality and inter-modality matching approaches have been introduced by many researchers. However, the proposed methods mostly ignore the effects of shape exaggeration of the facial sketch. In this paper, we attempt to investigate empirically and highlight the importance of considering the shape exaggeration effects in any of the two aforementioned matching approaches. This investigation is limited to the Chinese University of Hong Kong (CUHK) Face Sketch Database (CUFS) and CUHK Face Sketch FERET Database (CUFSF) datasets. The results demonstrate that the shape exaggeration effects may reduce retrieval rate accuracy.

Index Terms—Sketch to photo, identity of interest, shape exaggeration, forensic, feature extraction, intra-modality, inter-modality.

I. INTRODUCTION

A system that is able to identify Identity of Interest (IoI) from mugshot gallery based on its corresponding facial sketch has useful application in forensic investigation. This is to assist the law enforcement to short list some candidates before eventually apprehend the suspect. The facial sketch is called forensic sketch and it is usually rendered by a professional forensic artist like Lois Gibson [1] and Karen Taylor [2] when there is no other available evidence. Furthermore, the sketch is drawn solely based on the descriptions elicited from the eyewitness. This will cause the generated forensic sketch to have shape exaggeration with less accurate details.

Another problem that has been identified by Tang and Wang [3] is that the distance between sketch and photo of the same identity is larger than the distance between two photos of two different identity. This is due to a modality difference. There are two different approaches to address this. First, some researchers [3]–[15] attempt to reduce the modality gap at the pre-processing stage by synthesizing the image from one modality to the other modality before extracting and matching the feature vectors (i.e., intra-modality approach). Second, some other researchers [16]–[22] directly extract the feature vectors using modality-invariant descriptors and then perform the feature matching (i.e., inter-modality approach).

It is observed that most of the proposed methods in either intra-modality or inter-modality approaches do not cater for shape exaggeration effects. The problem seems to be at the feature extraction stage. For example, the intra-modality approach synthesize the gallery image into pseudo-photo or pseudo-sketch (i.e., pseudo-image) at the pre-processing stage before extracting the features. Despite the fact that the transformation algorithms are usually complex, the transformed pseudo-image is considered naive as it tries to exactly copy the shape of the image being transformed but in other modality. Figure 1 shows an example of the above mentioned observation. Without a sensitive shape exaggeration descriptors to extract the features, matching the facial sketch that contains shape exaggeration may result in a low retrieval rate accuracy although both images are in the same modality. Similar to that inter-modality approach where the shape exaggeration effects are typically not considered although the modality-invariant feature extractor has been used to represent the image. Overall, both approaches attempt to close the modality gap but the shape exaggeration effects are mostly ignored.

Recently, Difference of Gaussian Oriented Gradient Histogram (DoGOGH) has been demonstrated to be very effective in matching facial sketch to photo [23]. Next, this proposed method has been extended to Cascaded Static and Dynamic DoGOGH (C-DoGOGH) with intention to further improve the retrieval rate accuracy by catering the shape exaggerations effects [24]. It combines static and dynamic local feature extraction in a cascaded manner. In term of the efficacy of the proposed method, the Chinese University of Hong Kong (CUHK) Face Sketch Database (CUFS) and CUHK Face Sketch FERET Database (CUFSF) datasets were used in the experiment. Unfortunately, the retrieval rate accuracy improvement of the proposed method is only reported for inter-modality matching approach and never for intra-modality matching approaches. The fact that shape exaggeration effects will give accuracy impact for any of those approaches, therefore, this paper attempts to investigate and highlight the aforementioned effects regardless of the two matching approaches.

The rest of this paper is organized as follows. Section II outlines the existing methods on face sketch to photo matching. Section III elaborates the databases used and explain
Figure 1. Example images with shape exaggeration effects used in our study. Image (a) shows the sketch and (b) shows its corresponding photo. Image (b) has been synthesized using MWF [6] to produce a pseudo-sketch as shown in (c). It is noticeable that the generated pseudo-sketch in (c) naively follow the probe photo in (b). Note that the red patches with dotted line box indicate example patches that are observed to be misaligned due to the shape exaggeration of the sketch.

the experimental setup for this investigation. It is followed by Section IV that discusses the results obtained. Finally, the conclusion is drawn in Section V.

II. RELATED WORK

This research work has been pioneered by Uhl and Lobo [25]. To match forensic sketch to photos, the proposed method uses Eigenface and Principle Component Analysis (PCA). Later, Tang and Wang initiate a public dataset called CUFS [26] due to the fact that forensic sketches are usually confidential and not available for researchers. Next, another dataset called CUFSF has been introduced by Zhang et al. [27]. Unlike the CUFS dataset, the sketches in this dataset are closer to real forensic sketch because they were drawn with shape exaggeration. Based on these datasets, many state-of-the-art methods have been proposed by many researchers. From the literature, the proposed methods can be divided into two main approaches: intra-modality and inter-modality.

For the intra-modality approach, a synthetic image called pseudo-image is generated at the pre-processing stage. The aim is to ensure that the feature matching is performed within the same modality. This approach has been pioneered by Tang and Wang [3], [5]. It is followed by Gao et al. [4] and succeeding researchers [6]–[15]. Tang and Wang [3] transform a photo into sketch before feature matching using the proposed Eigensketch transformation algorithm. Gao et al. [4] proposed a synthesizing method based on Embedded Hidden Markov Models (E-HMM) to cater for the non-linear relationship between the sketch and photo. Later, Tang and Wang [5] synthesize the sketch or photo using the proposed synthesizing model based on Markov Random Fields (MRF). Peng et al. also utilized MRF to learn multiple representations of the image [9] and recently proposed a Superpixel-based synthesis method [10]. An improved MRF model called Markov Weight Fields (MWF) model has been proposed by Zhou et al. [6] to deal with patches selection and computational complexity. Due to a noisy synthesized sketch being generated, image denoising technique called Spatial Sketch Denoising (SSD) is exploited by Song et al. [8] to reduce the noise. To reduce the computational cost, recently, Wang et al. [12], [13] proposed the model-driven face sketch synthesis framework (LR) and offline random sampling for training the synthesizing model (RSLCR).

For the inter-modality approach, a modality-invariant features are often used to extract features from the images before matching [17]–[21]. It skips the synthesizing process at the pre-processing stage. Klare and Jain [16] proposed a method that extract the feature locally using a Scale Invariant Feature Transform (SIFT) descriptor. To improve the accuracy further, this method has been extended by fusing Multiscale Local Binary Pattern (MLBP) and the SIFT with Local Feature Discriminant Analysis (LFDA) [17]. Galoogahi and Sim [18] see that most of the research works use common features that are not meant for a cross-modality matching. Therefore a new face descriptor called Histogram of Averaged Oriented Gradients (HAOG) is proposed to extract modality-invariant features of salient facial components. The fact that facial shape is relatively invariant across modality, thus, Galoogahi and Sim [19] proposed a new face descriptor that is a shape-based and claimed to work on cross modality images. It is called Local Radon Binary Pattern (LRBP). The algorithm projects each non-overlapping patch on Radon space using Radon transform. Then, the features are extracted at every patch using Local Binary Pattern (LBP). Recently, Difference of Gaussian Oriented Gradient Histogram (DoGOGH) has been demonstrated to be very effective in matching facial sketch.
to photo [23]. To cater for shape exaggeration effects, this method has been extended to Cascaded Static and Dynamic DoGOGH (C-DoGOGH) with intention to further improve the retrieval rate accuracy by catering the shape exaggerations effects [24]. It combines static and dynamic local feature extraction in a cascaded fashion.

III. EXPERIMENTS

A. Databases

We use two different datasets in our experiments. The first dataset is CUHK Face Sketch Database (CUPS). This dataset was prepared by Tang and Wang [5], [26] and contains a total of 606 image pairs. The entire images are in frontal pose with neutral expression. The second dataset is CUHK Face Sketch FERET Database (CUFSF). This dataset was prepared by Tang and Wang [5], and Zhang et al. [27] and contains a total of 1194 image pairs. The sketches are drawn with shape exaggeration, and the photos are mostly exposed to lighting variation. In order to make a fair accuracy comparison between intra-modality and inter-modality matching approaches, we only use a total of 338 and 944 image pairs (similar to that of the number of generated pseudo-sketch) from CUPS and CUFSF datasets, respectively.

B. Experimental Setup

In order to empirically investigate the effect of shape exaggeration in the context of matching face sketch to photo, we need to compare the retrieval rate accuracy obtained from two different matching methods: insensitive and sensitive to shape exaggeration. For these, we employ Difference of Gaussian Oriented Gradient Histogram (DoGOGH) [23] and Cascaded Static and Dynamic DoGOGH (C-DoGOGH) [24], respectively. For the retrieval rate computation, we use Cumulative Match Curve (CMC) score. The rank-1 accuracy is compared between the DoGOGH and the C-DoGOGH matching methods. From the comparison, the shape exaggeration effects is observed if there is a rank-1 accuracy improvement.

The fact that we want to investigate the effect of shape exaggeration in both matching approaches (i.e., intra-modality and inter-modality), a general block diagram for the matching system is illustrated in Figure 2. We follow the same approach used in our previous publications (i.e., from [23] and [24]) where features are extracted locally from patch to patch as what has been initially proposed by Klare et al. [17] and the similarity measure is computed using nearest neighbour. The only different between the two matching approaches is at the pre-processing stage in which the path ‘1’ is used for intra-modality while the path ‘2’ is used for inter-modality (refer to Figure 2). The followings subsections elaborate more on each stage.

1) Pre-processing Stage: This stage first align the faces by using translation, rotation and scaling such that the angle between two eyes is 0 degrees with a certain pixels distance. Then, this image is cropped to size $250 \times 200$ with the eyes positioned at a pre-determined coordinates. Following the image size in [23] and [24], we resize the image to $175 \times 140$. For intra-modality approach, this stage executes the synthesizing algorithms to produce the pseudo-sketches as shown in Figure 2. In this work, the pseudo-sketches are generated by the following methods: RSLCR [13], Fast-RSLCR [13], LR [12], MRF [5], MWF [6] and SSD [8]. The example of the pseudo-sketches are shown in Figure 3. For inter-modality approach, this stage skips the transformation process and let the photos be as they are.

2) Feature Extraction Stage: We employed two different feature extraction methods that are DoGOGH [23] and C-DoGOGH [23]. The features are extracted locally from smaller overlapping patches of size $16 \times 16$. All these features are concatenated to built a full feature vector. We do not have
intention to elaborate the extraction methods further as all the
details can be obtained from the respective literature.

3) Similarity Measure Stage: Here, we consider untrained
classifier to compute the similarity that is based on nearest
neighbour. This is due to the fact that we want to utilize
all available samples that are very limited. Maximizing the
testing sample seems to be more appropriate in order to obtain
more objective accuracy. For that, we employ \( L_1 \)-distance as
in Equation (1) [28].

\[
D(x, y) = \sum_{i=1}^{L} |(x_i - y_i)|
\] (1)

Here, \( x \) and \( y \) represent the feature vectors from sketch
and photo, respectively. The feature vector length is denoted
as \( L \) and the distance is denoted as \( D \). From the computed
distances, we then compute the retrieval rate accuracy based
on the CMC.

IV. RESULTS

Table I and Table II show rank-1 accuracy comparison
evaluated on CUFS and CUFSF datasets, respectively. The
comparison was made between two different matching meth-
ods (i.e., DoGOGH and C-DoGOGH) across two different
matching approaches (i.e., inter-modality and intra-modality).
For intra-modality matching approach, some six different
pseudo-sketches are evaluated. Here, the CUFS dataset is con-
sidered as a clean dataset because the sketches were rendered
without shape exaggeration while the CUFSF dataset is the
dataset of interest for this investigation where the sketches
were rendered with shape exaggeration [5]. The investigation
here is based on the fact that if the C-DOGOGH can improve
the accuracy further than what is given by DoGOGH, therefore
the shape exaggeration is proven to be existed and thus reduce
the accuracy.

From the results in the tables, referring to the inter-
modality approach, C-DoGOGH improve the rank-1 accuracy
of DoGOGH by 0.30\% and 11.75\% on CUFS and CUFSF

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Inter-Modality (%)</th>
<th>Intra-Modality (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DoGOGH [23]</td>
<td>96.45</td>
<td>97.34</td>
</tr>
<tr>
<td>C-DoGOGH [24]</td>
<td>96.74</td>
<td>98.22</td>
</tr>
<tr>
<td>Improvement</td>
<td>0.30</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Inter-Modality (%)</th>
<th>Intra-Modality (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DoGOGH [23]</td>
<td>54.13</td>
<td>65.25</td>
</tr>
<tr>
<td>C-DoGOGH [24]</td>
<td>60.49</td>
<td>69.07</td>
</tr>
<tr>
<td>Improvement</td>
<td>11.75</td>
<td>5.85</td>
</tr>
</tbody>
</table>
datasets, respectively. This improvement behaviour is similar to what has been reported in [24] to claim that the shape exaggeration does effect the rank-1 accuracy. To investigate the same effect further on the intra-modality approach, all the rank-1 accuracy of the six pseudo-sketches are recorded in the Table I and Table II. Based on the results, the C-DoGOGH also improve the rank-1 accuracy regardless of any pseudo-sketches to replace the photos. Therefore, it indicates that the shape exaggeration is preserved in the pseudo-sketch because the fact that the synthesized sketch is naively follow the probe photo.

Of all rank-1 accuracy on CUFS dataset, matching sketch to pseudo-sketch that is generated by LR [12] give the highest rank-1 accuracy of 98.52% which is higher than the inter-modality matching approach. On the other dataset (i.e., CUFSF) the pseudo-sketch that is generated by RSLCR [13] give the highest rank-1 accuracy of 69.07% which is also higher than the inter-modality matching approach. From these observation, we see that the accuracy for intra-modality approach is relying on the quality of the generated pseudo-sketch. Higher the pseudo-sketch quality will results in better retrieval rate accuracy. Here, matching sketch to pseudo-sketch from RSLCR [13] shows better rank-1 accuracy across datasets.

Overall, it is obvious that the rank-1 accuracy improvement for dataset with shape exaggeration is larger than the dataset without shape exaggeration (i.e., clean dataset) in both matching approaches. In term of the highest rank-1 accuracy, the results demonstrate that the intra-modality matching approach gives better accuracy in comparison to the inter-modality matching approach provided that the pseudo-sketches are synthesized effectively with good quality. The retrieval rate comparison is also made for the first ten ranks. The results...
evaluated on CUFSS and CUFSSF datasets are shown in Figure 4 and Figure 5, respectively. From here, it can be seen that the retrieval rate accuracy can be slightly improved if the shape exaggeration effects is catered at the feature extraction stage as what has been proposed in C-DoGOGH [24].

V. CONCLUSION

In this paper, we presented an empirical investigation of the effect of shape exaggeration in the context of matching face sketch to photo. The results indicate that the C-DoGOGH (i.e., a matching method that is sensitive to shape exaggeration) manage to give better accuracy in comparison to the DoGOGH across two different matching approaches (i.e., intra-modality and inter-modality matching approaches). It is also observed that the rank-1 accuracy improvement for dataset with shape exaggeration (i.e., CUFSSF) is larger than the rank-1 accuracy improvement for clean dataset like CUFSS. Overall, regardless of either intra-modality or inter-modality matching approach, the results demonstrate that the shape exaggeration does effects the retrieval rate in any of the two matching approaches. This is due to the fact that the pseudo-sketch is synthesized naivefly to follow the probe photo.

ACKNOWLEDGMENT

This research work was supported in part by the Universiti Sains Malaysia Research University Individual (RUI) Research Grant Scheme no. 1001/PELECT/8014056, and the Universiti Teknologi MARA, Cawangan Pulau Pinang.

REFERENCES