IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 23, 2018, accepted June 11, 2018, date of publication July 12, 2018, date of current version August 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2855208

Difference of Gaussian Oriented Gradient Histogram for Face Sketch to Photo Matching

SAMSUL SETUMIN[1,2] AND SHAHREL AZMIN SUANDI[1], (Senior Member, IEEE)
[1]Intelligent Biometric Group, School of Electrical and Electronics Engineering, Universiti Sains Malaysia, Engineering Campus, Nibong Tebal 14300, Malaysia
[2]Faculty of Electrical Engineering, UniversitiTeknologi MARA Pulau Pinang, Permatang Pauh 13500, Malaysia

Corresponding author: Shahrel Azmin Suandi (shahrel@usm.my)

This work was supported in part by the Universiti Sains Malaysia Research University Individual Research Grant Scheme under Grant 1001/PELECT/814208 and in part by the Universiti Teknologi MARA.

ABSTRACT Retrieving a corresponding photo from a forensic sketch remains a great challenge. One of the difficulties is that the retrieval rate reduces significantly on data sets with shape exaggeration and lighting variation. The goal of this paper is twofold. First, we attempt to reduce the factor of the shape exaggeration problem by introducing new fiducial points for geometric face alignment. Next, to minimize the illumination effects, we describe the image using the difference of Gaussian-oriented gradient histogram. The matching results using the simplest distance measure on two public databases indicate that our proposed face sketch to photo matching method achieves significantly better accuracy than the state-of-the-art methods. It is also proved that shape exaggeration influences can be reduced by employing outer points face alignment.

INDEX TERMS Difference of Gaussian, identity of interest, image matching, oriented gradient histogram, sketch to photo.

I. INTRODUCTION

Seeking the right Identity of Interest (IOI) when there is no other evidence but a face sketch is continuously attracting researcher attention. In crime investigation, listing the identity of all relevant suspects based solely on descriptions given by an eyewitness is extremely challenging [1], [2]. A better way to do this is by rendering a face sketch based on the elicited descriptions and then match the sketch to photos in the mugshot database.

Despite this, it is not an absolute solution because the matching process is fairly hard and it is considerably complicated due to the fact that both images are from different modalities. To address this, some researchers have attempted to close this modality gap by converting a sketch to a pseudo-photo (or vice versa) so that both are in the same modality (i.e., intra-modality matching) [3]–[7]. The other researchers used modality-invariant features to represent the images and perform the similarity measure based on this representation (i.e., inter-modality matching) [8], [9]. For the latter approach, apart from having modality-invariant features to represent the image, the sketch and photo quality must also be taken into account because it may degrade the retrieval rate. To elaborate further, a sketch is drawn with no consideration of lighting conditions (i.e., no illumination) but it may suffer from slight shape exaggeration (especially for forensic sketches). While for photos, there is no possibility of shape exaggeration occurring, but there is potential of being exposed to lighting variations. Disregarding these imperfections will obviously sacrifice performance.

Also, due to the fact that sketches are drawn with no regard to the lighting conditions, matching the features from such representations is inaccurate. If mugshot photos are free from illumination variance, a better retrieval rate is expected because the extracted features are absent any illumination effects. Difference of Gaussians (DoG) is a well-known method for edge detection. As a face sketch is usually generated by emphasizing the edges of the facial components (i.e., the shape) more than the spatial regions, DoG seems to have potential to be utilized in that it increases the visibility of the edges of a face image regardless of lighting variations. As the shape appearance is obvious, a shape descriptor like Histogram of Oriented Gradient (HOG) can be employed to extract the features. Consequently, a new feature descriptor called Difference of Gaussian Oriented Gradient Histogram (DoGOGH) can be introduced to better represent the image regardless of the illumination.

Another aforementioned problem is the shape exaggeration effects. Note that the largest shape of a face is the face outline. Aligning the face using inner points like the center of the eyes will make the outer region be misaligned even worse. Klare *et al.* [8] reported that the outer regions of forensic sketches like hairline and chin are more salient than

the inner regions like eyes, nose, and mouth. Motivated by this finding, if we can assume that faces are geometrically aligned (i.e., at the preprocessing stage) using new reference points at the outer regions, it will help to reduce the effects of shape exaggeration because only the shapes of the inner facial components are exaggerated. Hence, the effect is not significant.

This paper is organized as follows. The related work is discussed in Section II. Then, the proposed method is explained in Section III. Section IV outlines the experimental procedures and gives a detailed discussion of the performances. Finally, Section V concludes the results.

II. RELATED WORK

The fact that face sketches and face photos are from different modalities, based on the literature, the proposed methods mostly fall under these two approaches: intra-modality and inter-modality. Any one of the approaches share the same objective, which is to get the best rank-1 accuracy possible. A Cumulative Match Curve (CMC) is a common evaluation tool used by most researchers in this field to compute the ranks accuracies [8], [10]–[19]. It measures the percentage of correct identity cumulatively across the ranks. As for an extreme example, if rank-1 percentage is at 100%, it simply means that the algorithm is able to recognize the face without mistakes being made. Likewise, if the cumulative percentage starts lower and gradually increases to reach 100% at rank-10, it indicates that the face candidate resides in the top 10 matches.

In the first approach, a synthetic image is generated at a preprocessing stage prior to a matching process. Most of the work in this approach was proposed by Tang and Wang [3], [4], Liu *et al.* [5], Wang and Tang [6], and Zhang *et al.*[7] which was then followed by Gao *et al.* [20], Wang *et al.* [21], [22], and Radman and Suandi [23] and succeeding researchers. The research works particularly focus on the viewed sketch database. In terms of performance, the state-of-the-art approach has achieved more than a 99% retrieval rate at the first rank as tested on the CUFS database (as reported by [24] and [25]). In the second approach, a direct matching process is employed on a common representation of a face sketch and photo. This technique is distinct from the intra-modality approach because there is no synthesis procedure execution. It extracts discriminative features that are invariant to photo and sketch modalities before performing a similarity measurement [8], [9], [11], [26], [27]. By doing this, the complex conversion process is eliminated but still demonstrates comparable accuracy.

Most of the research under the second approach utilizes local features extraction as in [8] to describe the image. It divides the image equally into patches, and extracts the features from each patch (i.e., local features). These local features are then concatenated to build a full feature vector that represents the image. Our proposed method is based on this approach. Note that features are extracted on a patch-by-patch basis. Misaligned patches (i.e., between sketch and photo) will definitely give wrong representations of the respective patch. Aligning the face using inner points (i.e., two eye centers [8], [10]; between eye centers and mouth center [27]; two eye centers and the mouth center [20], [24]) will make the outer region misaligned even worse. Klare *et al.* [8] reported that the outer regions of forensic sketches are more salient than the inner regions. To reduce this influence, we introduce new reference points that are picked at the outer regions for face alignment. Likewise, illumination effects are not properly catered for by most researchers. In order to compensate for this, we describe a face image using an illumination-invariant shape descriptor called Difference of Gaussian Oriented Gradient Histogram (DoGOGH).

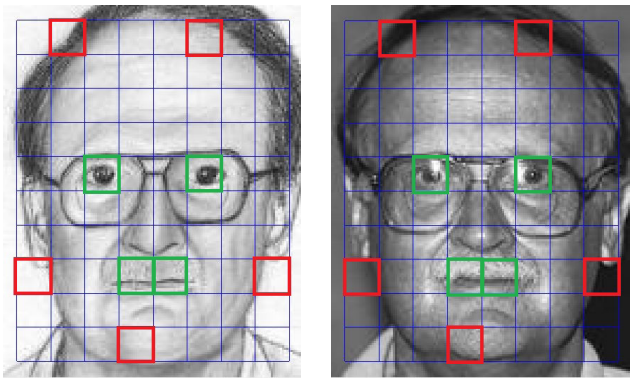
III. PROPOSED METHOD

Local feature descriptors have been successfully applied to give good accuracy in the context of matching sketches to photos [8]. They extract features locally from all patches and concatenates them to build a long feature vector. Here, we propose new fiducial points for face alignment so that the image is aligned using three fiducial points instead of two (i.e., as proposed by most researchers). Then, we extract the features using Difference of Gaussian Oriented Gradient Histogram (DoGOGH) to describe the face. The details of each process are elaborated in the following subsections.

A. NEW FIDUCIAL POINTS FOR FACE ALIGNMENT

Note that all face images come with various geometrical positions, orientations and sizes. Directly matching these images will result in poor matching accuracy. To improve this, all sketch and photo images need to be geometrically aligned such that the fiducial points (i.e., two eye centers [8], [10]; between eye centers and mouth center [27]; two eye centers and the mouth center [20], [24]) of all the face images fit into fixed reference points. In two fiducial points alignment, all the photos and sketches are aligned using image translation, rotation and scaling. For more than two fiducial points, affine transformation is employed. Doing face alignment will position similar face components from different images in roughly the same region.

The alignment process is very crucial because it may cause the computed similarity between two images to become lower due to misalignment and not due to the fact that images appear differently [10]. Klare *et al.* [8] reported that the outer regions of forensic sketches like hairline and chin are more salient than the inner regions like eyes, nose, and mouth. Aligning faces based on eyes and mouth points will make the outer patches (as shown in Fig. 1) somehow misaligned (especially for a sketch that has slight face shape exaggeration). Motivated by these findings, we propose new fiducial points for face alignment that are located at the outer regions of a face. This ensures that the patches at the outer region are matched correctly to their corresponding features in the matching process. This is because these regions carry salient features and therefore it is more discriminative than the inner regions. This is demonstrated in Section IV-B in this paper.



(a) (b)

FIGURE 1. Example images (CUHK Face Sketch FERET Database (CUFSF)) with facial shape exaggeration used in our study. Image (a) shows the sketch and (b) shows its corresponding photo. Green patches indicate example patches that are correctly aligned while red patches indicate example patches that are misaligned.

We select three fiducial points in this work. The points (as shown in Fig. 2 first left column) are at the left and right face edge (i.e., at the horizontal line of the eyes) and chin tip. By using affine transformation on these three points, the images are eventually cropped to size 175 × 140 with the fiducial points transformed to fixed reference points (as shown in Fig. 2 second left column).

B. DIFFERENCE OF GAUSSIAN ORIENTED GRADIENT HISTOGRAM

Here we provide a description of the proposed Difference of Gaussian Oriented Gradient Histogram (DoGOGH). Let $I(x, y)$ be the aligned image. We first convert this image into grayscale. Then, we apply gamma intensity correction $\hat{I}(x, y) = \log(I(x, y))$ to it to lighten the dark regions. It gives some level of robustness to lighting variations. Next, to obtain the DoGOGH features, we compute the Histogram of Oriented Gradient (HOG) on a Difference of Gaussian (DoG) image. The DoG image is free from both high-frequency and low-frequency illumination variations. The DoG image is computed as follows. The image $\hat{I}(x, y)$ is convolved with a Gaussian kernel $G_\sigma(x, y)$ as in (1) using two different width σ_1 and σ_2. Subtracting the convolved images as in (2) will end up getting the DoG image. It is also illustrated in Fig. 2.

$$G_\sigma(x, y) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

$$DoG(x, y) = G_{\sigma_1}(x, y) * \hat{I}(x, y) - G_{\sigma_2}(x, y) * \hat{I}(x, y) \tag{2}$$

Once the DoG image is ready, we extract the features using the HOG descriptor introduced by Dalal and Triggs [28]. This descriptor is designed for human detection but has been proven to work well in face recognition [29]. The descriptor is computed based on gradient vectors for each pixel in the DoG image. The following equation formulates the gradient vectors:

$$\nabla DoG(x, y) = \begin{bmatrix} G_x(x, y) \\ G_y(x, y) \end{bmatrix} = \begin{bmatrix} \frac{\partial DoG(x, y)}{\partial x} \\ \frac{\partial DoG(x, y)}{\partial y} \end{bmatrix} \tag{3}$$

These gradient vectors $\nabla DoG(x, y)$ are used to further compute the gradient orientation $\theta DoG(x, y)$ and magnitude $|\nabla DoG(x, y)|$ of each pixel in the DoG image

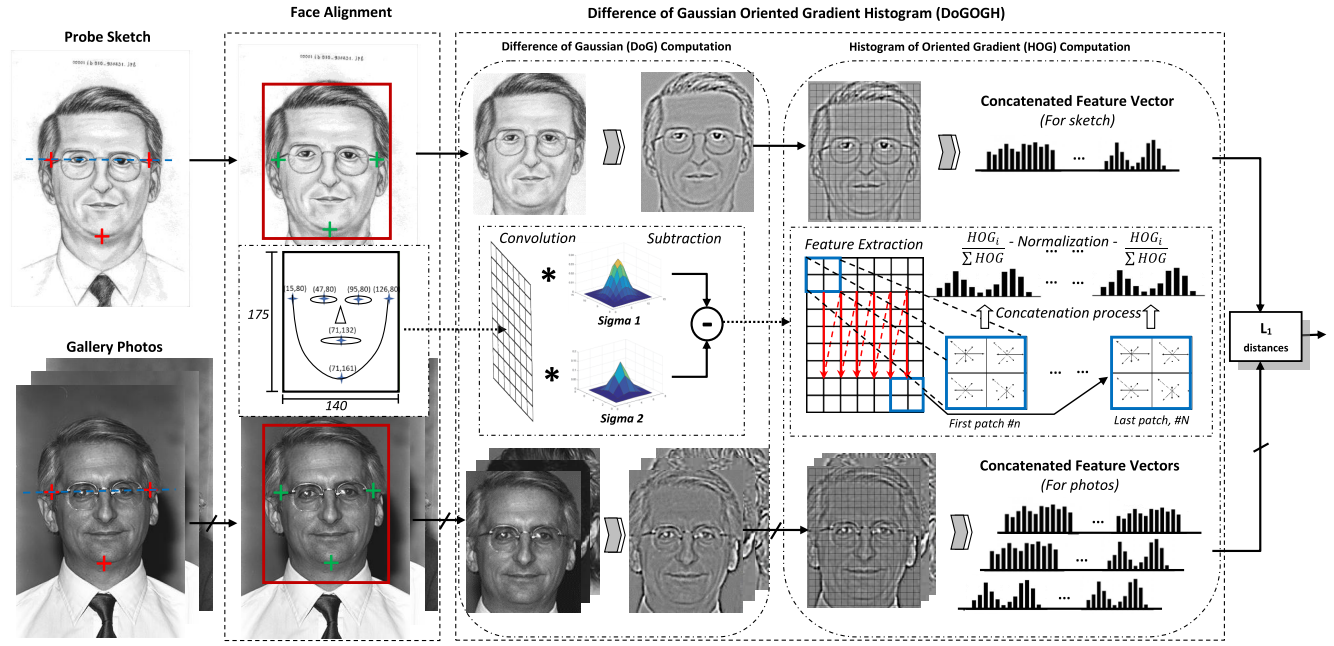


FIGURE 2. The proposed method for face sketch-to-photo matching. The method attempts to address problems with regard to shape exaggeration and illumination effects.

using (4) and (5), respectively.

$$\theta DoG(x,y) = tan^{-1}\frac{G_y(x,y)}{G_x(x,y)} \quad \in [-\pi,\pi] \tag{4}$$

$$|\bigtriangledown DoG(x,y)| = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{5}$$

To extract the features, we first divide the DoG image into small overlapping patches of size $N \times N$. Then, we extract the feature from patch to patch. On each patch, we compute a histogram in evenly spaced bins ranging from -π to π (i.e., to cater light-to-dark and dark-to-light transitions). The bin represents a certain orientation ranges, β, that is spaced based on the number of allocated bin, α (i.e., $\beta_b = \frac{\pi-(-\pi)}{\alpha}$, where $b = 1,\ldots,\alpha$). Every bin accumulates the gradient magnitude $|\bigtriangledown DoG(x,y)|$ of all pixel positions within a specified range (i.e., cell) according to its corresponding orientation $\theta DoG(x,y)$ that falls within β_b. The patch is divided into 2×2 cells. Each cell will produce an oriented gradient histogram. Concatenating these histograms will make up a HOG feature vector of the current patch, f_a. Here, we re-evaluate all four block normalization schemes (i.e., L_2-norm, L_2-Hys, L_1-norm and L_1-sqrt) evaluated by Dalal and Triggs [28] to select the best scheme in the context of matching sketches to photos. The L_1-norm and L_2-norm equations are defined in (6) and (7), respectively,

$$f'_{a(L1)} = \frac{f_a}{\|f_a\|_1 + \epsilon} \tag{6}$$

$$f'_{a(L2)} = \frac{f_a}{\sqrt{\|f_a\|_2^2 + \epsilon^2}} \tag{7}$$

where ϵ is a small constant. The extraction process is repeated for all patches across the image. We then concatenate these feature vectors to build a DoGOGH descriptor. Algorithm 1 shows the extraction details and Fig.2 illustrates the proposed method.

Algorithm 1 DoGOGH Feature Extraction Method

Input: Aligned face image $I(x,y)$.

***Step* 1**: *Preprocessing*. Convert the image into grayscale. Then apply gamma intensity correction $\hat{I}(x,y) = \log(I(x,y))$.

***Step* 2**: *DoG Image*. Compute $DoG(x,y)$ image from the preprocessed image $\hat{I}(x,y)$ using (2).

***Step* 3**: *Orientation and Magnitude Computation*. Compute orientation $\theta DoG(x,y)$ and magnitude $|\bigtriangledown DoG(x,y)|$ of each pixel on the $DoG(x,y)$ image using (4) and (5), respectively.

***Step* 4**: *Extract Features*. Divide the $DoG(x,y)$ image together with its orientation $\theta DoG(x,y)$ and magnitude $|\bigtriangledown DoG(x,y)|$ into small overlapping patches of size $N \times N$. Let $P = [p_a,\ldots,p_M]$ be the patches where $a = 1,2,\ldots,M$ and M is the total number of patches.

for each: $p_a \in P$

1: Initialize $f_a = []$.
2: $f_a \leftarrow |\bigtriangledown DoG_a|$ according to θDoG_a.
3: Normalize the f_a using L_2-norm, L_2-Hys, L_1-norm or L_1-sqrt to be f'_a.

Concatenate these feature vectors f'_a to build a DoGOGH descriptor $F = [f'_a,\ldots,f'_M]$.

Output: F.

Algorithm 2 Matching Algorithm

Input: Sketch feature vector F^S, Photo feature vector F_g^P where $g = 1,2,\ldots,G$. G is the size of gallery.

***Step* 1**: Calculate the L_1-distance d_g between F^S and F_g^P as follows:

$$d_g = \|F^S - F_g^P\|_1 \tag{8}$$

***Step* 2**: Sort d_g in ascending order. Let d_{g_s} be the sorted distance where g_s is the sorted indexes.

Output: The sorted indexes, g_s.

C. SIMILARITY MEASURE

In order to match a sketch feature vector F^S to photo feature vectors F_g^P, we use nearest neighbor (i.e., based on smallest distance) for classification. We conducted an experiment (empirically) to select an appropriate distance metric for this purpose. The results reveal that L_1-distance metric is the best similarity measure that is able to work reasonably well across several gradient-based descriptors. Hence, we employed the similarity measure as in Algorithm 2.

IV. EXPERIMENTS

In this section, we evaluate our proposed method on two public baseline datasets. The two datasets used are CUHK Face Sketch Database (CUFS) and CUHK Face Sketch FERET Database (CUFSF). Both datasets are categorized as a *Viewed Sketch* in which the forensic artists sketch the face while viewing the photo or the person being sketched.

The CUFS dataset was prepared by [3] and [6]. It contains 606 *Viewed Sketch* pairs from CUHK student dataset [30] (188 image pairs), AR dataset [31] (123 image pairs) and XM2VTS dataset [32] (295 image pairs). All the photos were in frontal pose, under normal lighting conditions, and with a neutral expression. For this dataset, only 311 (CUHK+AR) image pairs were available for testing in our experiments. Another dataset named CUFSF [6], [24] was also a *Viewed Sketch* drawn based on 1,194 photos from the FERET database [33]. The sketches were sketched with shape exaggeration with most of the photos exposed to lighting variation. Fig. 3 shows the example sketches with their corresponding photos. Since our proposed method does not require training, we used all available samples for testing.

A. EXPERIMENTAL SETUP

In sketch modality, the obvious dissimilarity between the sketch and its corresponding photo is the shape exaggeration. Although the sketches and photos have been aligned so that the fiducial points (e.g., center of eyes) are positioned at some fixed reference points, the facial shape from a sketch does not fit its correspondence well (sketch is rendered with slight shape exaggeration). Based on our observation, in this context

FIGURE 3. Examples of image pairs from (a)-(b) CUFS database (CUHK and AR respectively), and (c) CUFSF database; used in our evaluation.

we see that the face shape carries more discriminative features as compared to the other face components. This is because the region of face shape is larger than the other components and hence more discriminative shape features can be extracted from these regions. On the other hand, in photo modality, there is no significant shape exaggeration on it but it may be exposed to lighting variation (i.e., illumination effects). This illumination effect should not be an issue if we only consider a clean mugshot, for instance, photos from passports. But considering real-world application where the ideal photo does not always exist (e.g., CCTV images may be the reference), an illumination-invariant descriptor is required to handle such case.

Based on the aforementioned facts, we set up three experiments to evaluate the effectiveness of our proposed methods. The first experiment is to compare the rank-1 accuracy on different aligned images across some popular local descriptors. From this experiment, we expect to see that our newly introduced fiducial points give better results regardless of any local descriptors. This is to prove that the influence of shape exaggeration can be reduced by aligning the faces using fiducial points from the outer region (reported as more salient in [8]). The second experiment is to select the best block normalization schemes. The selection is based on which scheme gives the highest rank-1 accuracy. We evaluate L_2-norm, L_2-Hys, L_1-norm and L_1-sqrt as evaluated by Dalal and Triggs [28] in their work. Here, we expect to see which block normalization scheme normalizes the sketch and photo better as indicated by its matching accuracy. For our third experiment, we evaluate the performance (i.e., in terms of accuracy) of our proposed method in comparison with several popular local descriptors (used in the first experiment). We aim to see the performance for the first ten ranks. Here, we use our proposed fiducial points and the best block normalization scheme resulting from the second experiment. Finally, a performance comparison of the proposed method to the state-of-the-art methods is made.

In our experiments, all the images are aligned and cropped to size 175×140 and the fiducial points are transformed to a fixed reference points, which are left and right face edge, (r_1 and r_2) and chin tip (r_3), such that $\mathbf{r} = [r_1, r_2, r_3] = [(15, 80), (126, 80), (71, 161)]$. For DoG image computation, the two different widths used in the Gaussian kernel $G_\sigma(x, y)$ are $(\sigma_1, \sigma_2) = (1, 2)$ [24]. To extract the features, we use 16×16 (i.e., $N = 16$) 50% overlapping patch. Therefore, the features are extracted from 320 patches per image. For HOG computation, the number of bin, α, is set to 18.

The other four local descriptors, i.e., multiscale local binary patterns (MLBP) [8], scale-invariant feature transform (SIFT) [34], speeded up robust features (SURF) [35], and histograms of oriented gradients (HOG) [28], are used in this paper. Similarly, each local descriptor is extracted from image patches with size of 16×16. For SURF and HOG, we employ the implementation embedded in the MATLAB software. The SIFT feature vector is computed by exploiting an open source library [36]. All other settings and parameters used in our experiments are elaborated in the following sub-section. Note that the experiments are conducted using MATLAB R2016b under Windows 10 Pro 64 with 3.6GHz quad-cores processor and 16GB RAM.

B. THE EFFECT OF FACE ALIGNMENT

In the first experiment, we aimed to see the effect of face alignment in terms of its performance. To do the alignment, there are three common methods used by researchers. The first method is by performing translation, rotation and scaling so that the angle between two eyes is 0 degrees and the distance between the two is d_h pixels (i.e., 75 pixels) [8], [10]. Then, this image is cropped to size $H \times W$ (i.e., 250×200) with the eyes are vertically and horizontally centered at a

TABLE 1. Rank-1 accuracy comparison on the CUFS database. Five different local descriptors and four different fiducial points for alignment are evaluated.

Descriptors	Two Points (%)			Three Points (%)		
	Conventional HA	Proposed HA	*Improvement*	Conventional HAVA	Proposed HAVA	*Improvement*
MLBP	92.93	91.96	None	93.57	92.93	None
SIFT	91.96	92.28	0.35	91.00	92.28	1.41
SURF	91.00	91.32	0.35	91.32	89.71	None
HOG	98.39	97.75	None	98.07	99.36	1.32
DoGOGH	99.36	99.04	None	99.36	**100**	0.64

TABLE 2. Rank-1 accuracy comparison on the CUFSF database. Five different descriptors and four different fiducial points for alignment are evaluated.

Descriptors	Two Points (%)			Three Points (%)		
	Conventional HA	Proposed HA	*Improvement*	Conventional HAVA	Proposed HAVA	*Improvement*
MLBP	21.61	29.82	37.99	38.02	39.87	4.87
SIFT	23.79	35.18	47.88	43.13	48.83	13.22
SURF	15.16	23.28	53.56	40.70	47.91	17.71
HOG	31.16	47.91	53.75	64.24	72.11	12.25
DoGOGH	36.35	53.02	45.86	66.67	**83.75**	25.62

predetermined point (i.e., at row 115). We call this method as Horizontal Alignment (HA). The second method [27] is similar to that in the first method but with different fiducial points which are from the center between the eyes and center of the mouth. The distance d_v between the two is 78 pixels and positioned at column 100. We call this method as Vertical Alignment (VA). The third method combines these two to get three fiducial points for alignment. This method is used by [20] and [24]. We call this method as Horizontal and Vertical Alignment (HAVA). Note that all images from these three methods are centered at a common pixel point $P_{com} = (100, 115)$. Based on our experiment, aligning the faces using HAVA demonstrates slightly higher average accuracies on datasets with shape exaggeration (refer to Table 2). After alignment, we crop the image to size 175×140 and is centered at a common pixel point $P_{com} = (71, 80)$ (i.e., 70% smaller than 250×200 and $(100, 115)$, respectively). This is done because we want to reduce the feature dimension as well as its computational time.

The fact that a rendered sketch normally has some degrees of shape exaggeration (especially on forensic sketches) that make some parts of the face geometrically misaligned (as illustrated in Fig. 1) may result in a low recognition rate. If the feature vector construction is constructed based upon the right image patches, then the recognition rate could be increased due to the fact that the patch comparison is made up on the right pairs. Although all faces can be aligned using the aforementioned three fiducial points (i.e., HAVA), the outer regions that carry more discriminative features are not properly aligned. Hence, we proposed a new HAVA. Since, the selected fiducial points are at the outer regions, we manually annotate those points as defined in Subsection III-A. Fig. 4 shows the comparison between a commonly used face alignment (i.e., HA) and the proposed HAVA. It is clearly seen that the proposed HAVA aligns the face better (lesser white regions, where white region indicates the differences) as shown in the example in Fig. 4 (c) and (d). We analyzed one of the patches from this region by computing the L_1-distance (as shown in Fig. 5). The L_1-distance of the patch pairs aligned using the proposed HAVA is smaller than the L_1-distance of the patch pairs aligned using HA. Smaller L_1-distance simply means that the patch pairs have higher similarity. The results indicate that the similarity is higher when the patch is aligned using the proposed alignment. Also, we can visually see that the patch from the image that is aligned by the proposed HAVA has a more similar appearance as compared to the other one.

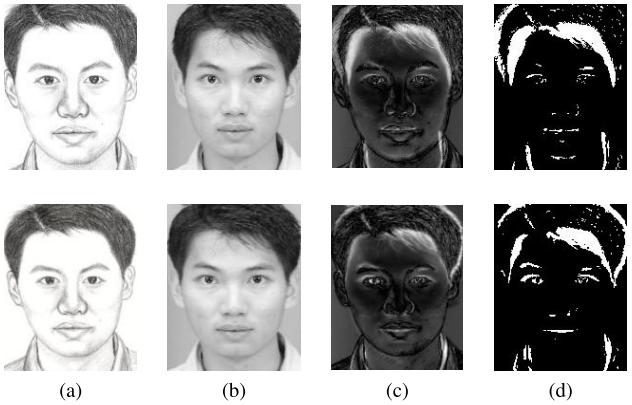


FIGURE 4. Example face that has been aligned using two different fiducial points. The first row is aligned using two fiducial points from the center of the eyes (i.e., HA). The second row is aligned using newly introduced fiducial points (i.e., proposed HAVA). Image (a) sketch (b) its corresponding photo, and (c) the difference between (a) and (b). Image (d) is the binarized image in (c) where the white region indicates the misalignment or exaggerated parts. The percentage of white pixel for HA is 13% while the proposed HAVA is 10.56%.

To further evaluate the effectiveness of the proposed HAVA, we compared the rank-1 accuracy on different aligned images (i.e., conventional HA: center of two eyes, proposed HA: left and right face edge, conventional HAVA: center of two eyes and mouth center, and proposed HAVA: left and right face edge and chin tip) across four popular local descriptors (i.e., MLBP, SIFT, SURF and HOG) and our proposed descriptor, DoGOGH. Vertical alignment, VA was excluded from this evaluation because its rank-1 accuracy is very similar to that of HA. Here, we use L_1-norm for the block normalization. The results in Table 1 demonstrate the capability of our proposed descriptor to achieve 100% accuracy when the face is aligned using our proposed HAVA. This is a clean dataset in which there is no significant shape exaggeration on the sketch. To prove that the influence of shape exaggeration can be reduced by aligning the faces using our proposed alignment points, we tabulated the results tested on the CUFSF dataset in Table 2. From the results, we can clearly see that three-point alignments have better accuracies as compared to two-point alignments. For the two-point alignment, using our proposed points (i.e., proposed HA) gives significantly higher accuracies than using the commonly used HA

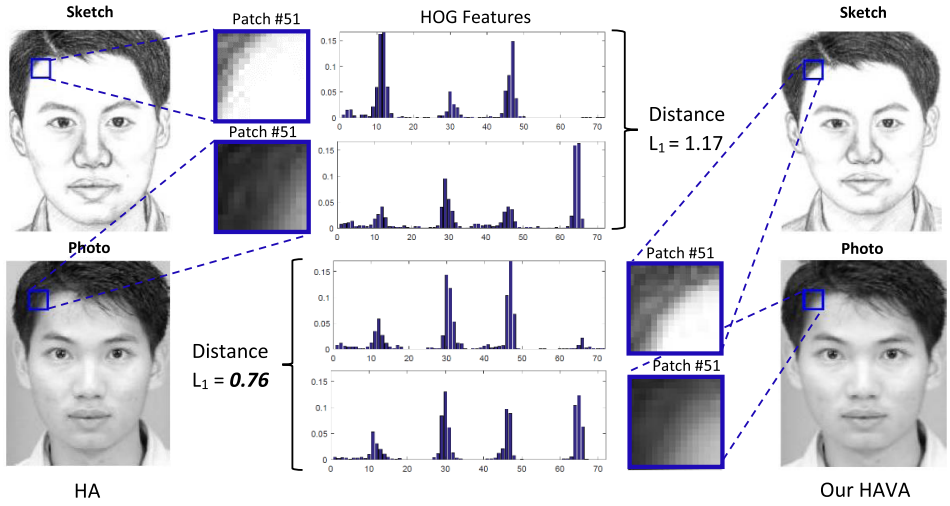


FIGURE 5. Comparison of L_1 distances between two different alignments of the same patch. Using our proposed HAVA give smaller L_1 distance as compared to the HA. Smaller L_1 distance simply means that the patch pairs have higher similarity. The selected patch is taken from the outer region on parts that have a larger difference when aligned using HA.

(i.e., conventional HA). This is obvious across descriptors. Similarly, for the three-point alignment, our proposed HAVA gives noticeably higher accuracies in comparison with the commonly used HAVA (i.e., conventional HAVA). This is also obvious across the local descriptors. Overall, three-point alignment shows better accuracy in comparison to two-point alignment and aligning the images using our proposed HAVA gives the best accuracies across descriptors. Note that we use L_1-distance for the similarity measure as described in Section III-C.

C. THE EFFECT OF BLOCK NORMALIZATION

The second experiment was conducted in order to select the best block normalization schemes. This is important for contrast normalization on each patch. The evaluation in Dalal and Triggs [28] reported that L_1-norm performs 5% lower than the performance of L_2-norm, L_2-hys and L_1-sqrt. But the evaluation is not in the context of matching sketches to photos. Therefore, we attempted to re-evaluate those normalization schemes in this context. The best normalization scheme was selected based on which scheme gave the highest rank-1 accuracy. The results are shown in Table 3. From the results, L_1-norm gives the highest accuracy of 100% and 83.75% on CUFS and CUFSF datasets, respectively. This contradicts the reported findings due to contextual differences. Note that the value of ϵ here is set to 0.01.

TABLE 3. Rank-1 accuracy comparison on the CUFS and CUFSF datasets using four different block normalization schemes with ϵ is set to 0.01.

Datasets	Block Normalization Schemes (%)			
	L_1-Norm	L_1-Sqrt	L_2-Norm	L_2-Hys
CUFS	**100**	99.04	99.36	99.04
CUFSF	**83.75**	80.65	78.89	80.23

D. FACE SKETCH TO PHOTO MATCHING

The recognition rate is considerably good for a clean dataset (i.e., frontal pose, under normal lighting conditions, and with a neutral expression) but poor on datasets with illumination variance (examples are shown in Fig. 6). This is because the sketches are drawn with no consideration of lighting conditions. To get a better performance, the images must be free from illumination effects or an illumination-invariant descriptor must be used to extract the feature vectors. Hence, the DoGOGH is proposed here to extract the features.

In the third experiment, the effectiveness of DoGOGH was evaluated and compared with several popular local descriptors (i.e., MLBP, SIFT, SURF, and HOG). This is to show that the proposed descriptor can give high accuracy when tested on a dataset without or with illumination effects. Considering the fact that face alignment is critical in the context of matching sketches and photos due to the shape exaggeration factor, all the images are first aligned using our proposed HAVA.

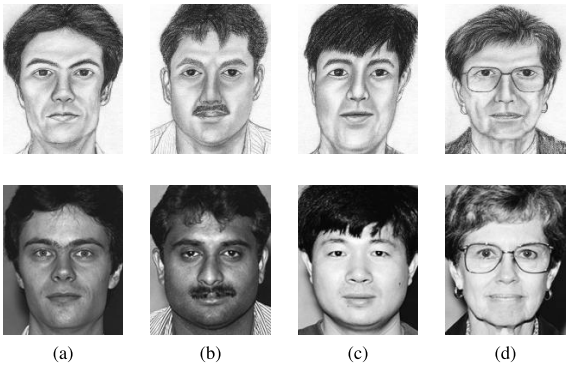
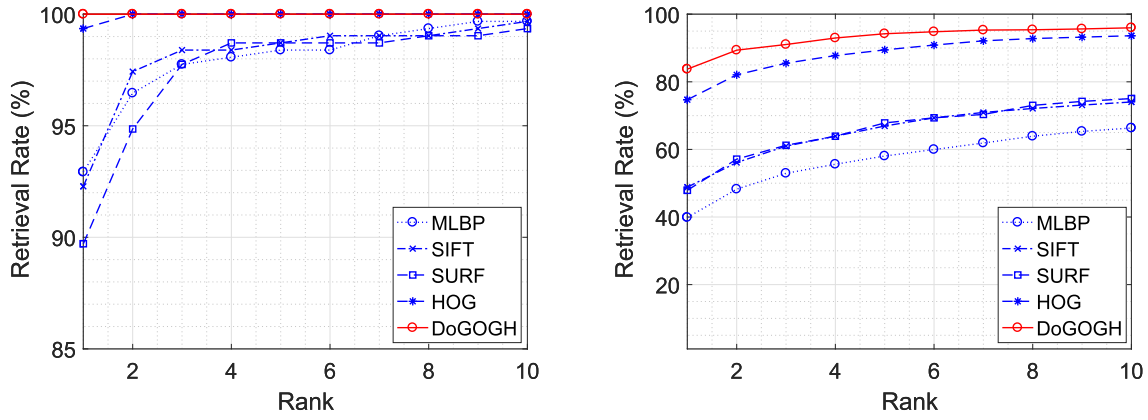


FIGURE 6. Example images with lighting variation used in our study. Image (a) to (d) are from one dataset. The top row shows the sketch and its corresponding photo is in the bottom row.



(a) CMC accuracies of the first ten ranks evaluated on CUFS.

(b) CMC accuracies of the first ten ranks evaluated on CUFSF.

FIGURE 7. Retrieval rate comparison of DoGOGH and several popular local descriptors evaluated on (a) CUFS and (b) CUFSF.

TABLE 4. Rank-1 CMC accuracy (%) of state-of-the-art methods on the CUFS and CUFSF datasets. Evaluation settings and accuracies are taken from the respective publications.

State-of-the-art methods	No. training samples	No. testing samples	Rank-1 accuracy (%)
CUHK Face Sketch Database (CUFS)			
MMRF + RS-LDA [6]	306	300	96.30
CITE [24]	306	300	99.87
SIFT + MLBP [8]	306	300	99.47
SNS-SRE [37]	306	300	96.50
TFSP + RS-LDA [22]	306	300	97.70
MrFSPS + RS-LDA [18]	306	300	97.70
S-FSPS + LDA [25]	306	300	99.10
DoGOGH	0	311	**100**
CUHK Face Sketch FERET Database (CUFSF)			
TFSP + RS-LDA [22]	500	694	72.62
MrFSPS + RS-LDA [18]	500	694	75.36
S-FSPS + LDA [25]	500	694	72.19
DoGOGH	0	1194	**83.75**

While extracting DoGOGH features, the extracted features are normalized using L_1-norm block normalization scheme. For matching, L_1-distance is used to measure the similarity. By using these settings, the accuracies were plotted across the first ten ranks as shown in Fig. 7. It demonstrates that DoGOGH performs better than all other local descriptors. It can easily achieve a 100% retrieval rate on CUFS (clean dataset) at rank-1, whereas on the CUFSF dataset, DoGOGH retrieves the faces at the rates of 83.75% and 95.98% for rank-1 and rank-10, respectively. This simply means that only 194 faces out of 1,194 faces were wrongly matched and only 48 faces were wrongly identified if the correct match was searched within the first 10 sorted candidates according to Algorithm 2. Despite that, Table 4 lists the performance comparison between the proposed method and the state-of-the-art methods. From Table 4, our proposed DoGOGH performs better than the other methods although no training phase is required. As our proposed method operates under the inter-modality approach, a comparison (in the table) was also made with the other *inter-modality* approaches. There are Coupled Information-Theoretic Encoding (CITE) [24] and Scale-Invariant Feature Transform (SIFT) + Multiscale Local Binary Pattern (MLBP) [8]. The results suggest that the *inter-modality* approach can be used to outperform *intra-modality* approach with less preprocessing complexity. Interestingly, both approaches have the same objective of achieving high matching accuracy.

V. CONCLUSION

In this paper, we propose DoGOGH as a new hand-crafted feature descriptor for sketch to photo matching. The proposed descriptor is designed such that it is immune to illumination effects. Overall, the matching accuracies using the simplest distance measure on two public databases (i.e., CUFS and CUFSF) indicate that DoGOGH achieves significantly better accuracy than the state-of-the-art methods. Furthermore, it is also proven that DoGOGH works well on datasets with illumination effects. These accuracies may be further improved by employing a better classifier, which exceeds the scope of this paper. In terms of the influence of shape exaggeration, it can be reduced by utilizing the new proposed fiducial points (i.e., proposed HAVA) for face alignment in the preprocessing stage. Using an inter-modality approach may reduce the preprocessing complexity

and hence speed up the matching time. Further analysis of how feasible this proposed method is on real forensic images is the path forward.

REFERENCES

[1] J. W. Shepherd, "An interactive computer system for retrieving faces," in *Aspects of Face Processing*. Dordrecht, The Netherlands: Springer, 1986, pp. 398–409.
[2] B. F. Klare, S. Klum, J. C. Klontz, E. Taborsky, T. Akgul, and A. K. Jain, "Suspect identification based on descriptive facial attributes," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep./Oct. 2014, pp. 1–8.
[3] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 687–694.
[4] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, Jan. 2004.
[5] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 1005–1010.
[6] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.
[7] W. Zhang, X. Wang, and X. Tang, "Lighting and pose robust face sketch synthesis," in *Computer Vision—ECCV* (Lecture Notes in Computer Science and Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6316. Berlin, Germany: Springer, 2010, pp. 420–433.
[8] B. F. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011.
[9] H. K. Galoogahi and T. Sim, "Inter-modality face sketch recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 224–229.
[10] B. Klare and A. K. Jain, "Sketch-to-photo matching: A feature-based approach," *Proc. SPIE*, vol. 7667, p. 766702, Apr. 2010.
[11] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.
[12] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "On matching sketches with digital face images," in *Proc. 4th IEEE Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2010, pp. 1–7.
[13] S. Klum, H. Han, A. K. Jain, and B. Klare, "Sketch based face recognition: Forensic vs. composite sketches," in *Proc. Int. Conf. Biometrics*, 2013, pp. 1–8.
[14] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, "The FaceSketchID system: Matching facial composites to mugshots," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2248–2263, Dec. 2014.
[15] P. Mittal, M. Vatsa, and R. Singh, "Composite sketch recognition via deep network—A transfer learning approach," in *Proc. Int. Conf. Biometrics*, 2015, pp. 251–256.
[16] S. Ouyang, T. Hospedales, Y. Z. Song, and X. Li, "Cross-modal face matching: Beyond viewed sketches," in *Computer Vision—ACCV* (Lecture Notes in Computer Science and Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9004. Cham, Switzerland: Springer, 2015, pp. 210–225.
[17] H. Roy and D. Bhattacharjee, "Face sketch-photo recognition using local gradient checksum: LGCS," *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 5, pp. 1457–1469, 2017.
[18] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, "Multiple representations-based face sketch–photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2201–2215, Nov. 2016.
[19] Z. Chen, K. Wang, and C. Liu, "Fast face sketch–photo image synthesis and recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 30, no. 10, pp. 1656008-1–1656008-13, 2016.
[20] X. Gao, J. Zhong, D. Tao, and X. Li, "Local face sketch synthesis learning," *Neurocomputing*, vol. 71, nos. 10–12, pp. 1921–1930, Jun. 2008.
[21] N. Wang, X. Gao, D. Tao, and X. Li, "Face sketch-photo synthesis under multi-dictionary sparse representation framework," in *Proc. 6th Int. Conf. Image Graph.*, Aug. 2011, pp. 82–87.
[22] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "Transductive face sketch-photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1364–1376, Sep. 2013.
[23] A. Radman and S. A. Suandi, "Robust face pseudo-sketch synthesis and recognition using morphological-arithmetic operations and HOG-PCA," *Multimedia Tools Appl.*, pp. 1–22, Feb. 2018.
[24] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 513–520.
[25] C. Peng, X. Gao, N. Wang, and J. Li, "Superpixel-based face sketch–photo synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 2, pp. 288–299, Feb. 2017.
[26] H. K. Galoogahi and T. Sim, "Face sketch recognition by local radon binary pattern: LRBP," in *Proc. Int. Conf. Image Process. (ICIP)*, 2012, pp. 1837–1840.
[27] M. A. A. Silva and G. C. Chávez, "Face sketch recognition from local features," in *Proc. Brazilian Symp. Comput. Graph. Image Process.*, 2014, pp. 57–64.
[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[29] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients," *Pattern Recognit. Lett.*, vol. 32, no. 12, pp. 1598–1603, 2011.
[30] X. Tang and X. Wang, "Face photo recognition using sketch," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2002, pp. 257–260.
[31] A. Martínez and R. Benavente, "The AR face database," CVC, Barcelona, Spain, Tech. Rep. #24, 1998.
[32] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authentication*, vol. 24, 1999, pp. 72–77.
[33] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[35] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.
[36] A. Vedaldi and B. Fulkerson. (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. Accessed: Nov. 2, 2017. [Online]. Available: http://www.vlfeat.org/
[37] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1213–1226, Aug. 2012.

SAMSUL SETUMIN received the B.Eng. degree (Hons.) in electronic engineering from the University of Surrey in 2006 and the M.Eng. degree in electrical, electronic and telecommunication from Universiti Teknologi Malaysia in 2009. He is currently pursuing the Ph.D. degree with Universiti Sains Malaysia. He was a Test Engineer with Agilent Technologies (M) Sdn. Bhd. and later he was with Intel Microelectronics (M) Sdn. Bhd. for a year. Since 2010, he has been a Lecturer with Universiti Teknologi MARA Pulau Pinang, Malaysia. His research interests include computer vision, image processing, and pattern recognition.

SHAHREL AZMIN SUANDI received the B.Eng. degree in electronic engineering in 1995 and the M.Eng. and D.Eng. degrees in information science from the Kyushu Institute of Technology, Fukuoka, Japan, in 2003 and 2006, respectively. He was an Engineer with Sony Video (M) Sdn. Bhd. and Technology Park Malaysia Corporation Sdn. Bhd. for almost six years. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, Universiti Sains Malaysia (USM), Engineering Campus, Penang, Malaysia. At USM, he serves as the Coordinator of the Intelligent Biometric Group, where he is currently the Founder of a biometric product, FaceBARSİ. His current research interests are face-based biometrics, real-time object detection and tracking, and pattern classification. He has served as a reviewer for several international conferences and journals, including the *IET Biometrics*, *IET Computer Vision*, *Multimedia Tools and Applications*, *Neural Computing and Applications*, *Journal of Electronic Imaging*, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and others.