IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 9, 2020, accepted June 15, 2020, date of publication July 16, 2020, date of current version August 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009744

Canonical Correlation Analysis Feature Fusion With Patch of Interest: A Dynamic Local Feature Matching for Face Sketch Image Retrieval

SAMSUL SETUMIN[1,2], (Member, IEEE), MUHAMAD FARIS CHE AMINUDIN[1], AND SHAHREL AZMIN SUANDI[1], (Senior Member, IEEE)

[1]Intelligent Biometric Group, School of Electrical and Electronics Engineering, Universiti Sains Malaysia, Nibong Tebal 14300, Malaysia
[2]Faculty of Electrical Engineering, Universiti Teknologi MARA, 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 (RUI) Research under Grant 1001/PELECT/8014056, and in part by the Universiti Teknologi MARA.

ABSTRACT An automatic photo retrieval system based on a face sketch has very useful application as to narrow down potential suspects in criminal investigations. This is true when there is no other evident except the face sketch that is rendered based on the recollection of a victim or eyewitness. Among the noticeable difficulties in matching the sketch and photo due to its modality difference are the generated sketch has some tendency of shape exaggeration, the sketch has very less accurate details and the real-world photo may expose to lighting variation unlike the sketch. In this paper, we attempt to address these complications by matching the sketch and photos using dynamic local feature of Difference of Gaussian Oriented Gradient Histogram (DoGOGH) on some selected patches. To avoid discriminative power degradation due to a large number of gallery images, two stage matching blocks are introduced in a cascaded fashion. The front block matches the feature such that it short lists k most similar photos for the second block. In this front block, Histogram of Oriented Gradient (HOG) and Gabor Wavelet (GW) features are fused by maximizing the correlation between the two using Canonical Correlation Analysis (CCA). Based on the short listed photos, the following block re-matched the sketch and photos using dynamically extracted local feature on its Patch of Interest (PoI). Eventually, the matching scores from the blocks are fused before getting rank-1 accuracy. The experimental results on two baseline datasets indicate that the proposed method outperforms the state-of-the-art methods. The extended evaluation on semi-forensic and forensic sketch datasets demonstrate its usage feasibility.

INDEX TERMS Identity of interest, patch of interest, sketch-to-photo, face sketch, forensic, image retrieval, CCA fusion, score fusion, deep learning.

I. INTRODUCTION

Face sketch image retrieval system is a system that capable of retrieving the corresponding photo from a face sketch. It is useful in criminal investigation to expedite the process of narrowing down potential suspects from a large mugshots. The system attempt to identify the Identity of Interest (IOI) based on a sketch when there is lack of other evidence at hand. This sketch is a hand drawn sketch that is sketched by forensic artist merely based on the descriptions elicited from the victim or eyewitness [1], [2]. The fact that the face sketch is drawn based on recollection of a victim or eyewitness, the generated sketch may not be accurate and has some degree of shape exaggeration. The minute details are also missing. Furthermore, the real-world mugshots may also expose to lighting variations. Retrieving the corresponding photo from such constraints are considerably hard because the fact that both images are from different modality. Thus, it

The associate editor coordinating the review of this manuscript and approving it for publication was Maurizio Tucci.

continuously attracts researchers attentions to propose a proper method to match these kind of images. Some researchers eliminate the modality difference by transforming the sketch to photo or vice versa such that the images are in the same modality before executing a matching process (i.e., intra-modality matching) [3]–[13]. The other researchers try to avoid the transformation complexity by directly represent the images using a common modality-invariant features and hence, the similarity measure is computed based on this representation (i.e., inter-modality matching) [14]–[19].

In the field of multimodal recognition, feature fusion based on Canonical Correlation Analysis (CCA) [20] has attracted researcher attention. The analysis attempts to find linear combinations of two different sets of feature vectors such that it maximizes the correlation between the two. This is done to increase the discriminative power. Based on this strategy, it becomes very popular and many CCA-based methods have been proposed in this research area [21]–[25]. Motivated by these findings, we adapt the same feature fusion approach to perform the matching. The fact that this approach does not really consider the effect of lighting variation, shape exaggeration and less minute details, therefore another matching block is introduced in this paper. The block is connected in a cascaded fashion to that CCA-based matching block with the aims to cater those aforementioned constraints. In the latter block, the Difference of Gaussian Oriented Gradient Histogram (DoGOGH) descriptor is used for feature extraction [26]. This is to address the illumination variance problem. To tackle imperfections due to shape exaggeration and less accurate details, a dynamic local feature matching is employed on some selected patches called the Patch of Interest (PoI). The PoI is obtained from an algorithm that is designed to reduce the number of patches by selecting only some meaningful patches instead of selecting the dense grid and uniform patches across the face image.

The main contribution of this paper is the proposed Patch of Interest (PoI). It is followed by the other two contributions that are the dynamic local feature extraction on the PoI and the cascaded matching blocks with score-level fusion. The contributions of this paper are therefore threefold. First, we cascade CCA-based feature fusion matching block (i.e., first stage) and second stage matching block for rank-1 accuracy improvement. The score from first stage and second stage are fused using score-level fusion. Secondly, we propose to match the sketch and photos only at the PoI such that it improves matching time and accuracy. Finally, we introduce a dynamic local feature on PoI matching method. To the best of our knowledge, no other local feature extraction method in the literature using this kind of approach.

In order to demonstrate the effectiveness of the proposed method, two baseline datasets (i.e., Chinese University of Hong Kong (CUHK) Face Sketch Database (CUFS) and CUHK Face Sketch FERET Database (CUFSF)) are used in the experiment. Although these datasets are classified as viewed sketch (i.e., drawn while viewing the photo), but the generated sketch in the CUFS dataset still has slight shape exaggeration while the sketch in CUFSF dataset has more shape exaggeration and thus are closer to real forensic sketches.

This paper is organized as follows. The related work is discussed in Section II. Then, the proposed method are explained in Section III and IV. Section V outlined the experiment procedures and give a details discussion upon the performances. Finally, Section VI concludes the results.

II. RELATED WORK

In law enforcement, the conventional method to narrow down potential perpetrators is by the onlooker to select a few candidates while browsing a large number of mugshot photos. It will obviously consume time. Apart from that, choose the candidates using this approach may disturb the recollection of the onlooker in the sense that too many browsed photographs are seen for quite some time (e.g., the onlooker experience memory fatigue). Thus, lessen the accuracy. Alternatively, an automatic photo retrieval system can be implemented to speed up the selection process without sacrificing the accuracy. To this end, a forensic sketch is required. The sketch is a sketch drawn (i.e., on paper using charcoal pencil) by forensic artist based merely on the descriptions elicited from the onlooker. Lois Gibson and Karen Taylor are of the well-known forensic artists who involve in this kind of sketch [2], [27]. While interviewing the onlooker or the victim, the artists comprehend the descriptions given and visualize it in their mind. Then the visualization is translated into a sketch by obeying a certain procedure as in [1]. Eventually, an electronic scanner is normally used to digitize the sketch before employing a matching algorithm.

This research work has been pioneered by Jr and Lobo [28]. They utilized Eigenface and Principle Component Analysis (PCA) to match the sketch and photos. Due to the fact that forensic sketch are always confidential, Tang and Wang took an initiative to create a public dataset named CUFS [3]. This is a viewed sketch (i.e., sketched while viewing the photograph) dataset created with very less shape exaggeration. Some time later, for research advancement, another dataset called CUFSF [29] is introduced by Zhang *et al.* This time, the sketches are generated with more shape exaggeration (so that it is closer to real forensic sketches) with the corresponding photo exposed to lighting variations. Based on these two baseline datasets, a lot of researchers continue proposing the new matching methods to get better state-of-the-art performance. The fact that face sketch and face photo are from different modality, it is known from the literature that most of the proposed methods are under these two main approaches: intra-modality and inter-modality.

In the intra-modality approach, a pseudo-sketch or pseudo-photo is generated at the preprocessing stage before a matching process. Most of the work in this approach were proposed by Tang and Wang [3], [4], Liu *et al.* [5], Wang and Tang [7], Zhang *et al.* [8]. It is then followed by Gao *et al.* [6] and succeeding

researchers [9], [10], [12], [13], [30]–[33]. Tang and Wang [4] transform a photo into sketch (i.e., pseudo-sketch) by using Eigensketch transformation algorithm before matching. Qingshan *et al.* [5] proposed a synthesizing technique based on Kernel-based Nonlinear Discriminant Analysis (KNDA) that preserves the local geometry. Gao *et al.* [6] used Embedded Hidden Markov Models (E-HMM) to model the nonlinear relationship between the sketch and photo while generating the pseudo-sketch. Later, Wang and Tang [7] proposed a synthesizing model based on Markov Random Fields (MRF) to produce pseudo-sketch or pseudo-photo. Zhang *et al.* improved this model to work under pose and lighting variations. Gao *et al.* [34] proposed Sparse Neighbor Selection (SNS) to render the initial pseudo-image and then used Sparse Representation-based Enhancement (SRE) to improve the quality of the synthesized image. Wang *et al.* [10] introduced probability graphic model and transductive learning to minimizes the empirical loss for training samples. Peng *et al.* employed MRF for multiple representations learning [11] and recently proposed Superpixel-based synthesis method [12] to synthesize the images. This was then followed by Cao *et al.* [32] who proposed Asymmetric Joint Learning (AJL) that attempts to cater for image discrepancies due to the modality difference. Other researchers have explored a deep learning approach to synthesize the image [30], [35], [36]. Overall, it should be noted that the synthesizing algorithms are often more complex than the recognition task itself [37].

For the inter-modality approach, it distinct to intra-modality approach in which it skips the synthesizing or conversion process. Generally, there are two main approaches to extract features: handcrafted feature [15]–[19] and deep learning feature [38]–[43]. Here, we attempt to contribute to the handcrafted feature approach. A common representation of the face sketch and photo is extracted and matched. It extracts discriminative features that are invariant to photo and sketch modalities before performing similarity measure [15]–[19], [26], [44]. Klare and Jain [14] proposed a local feature extraction approach using Scale Invariant Feature Transform (SIFT) descriptor. Later, this method is extended by fusing Multiscale Local Binary Pattern (MLBP) and SIFT with Local Feature Discriminant Analysis (LFDA) [15]. Zhang *et al.* proposed a new face descriptor based on Coupled Information-Theoretic Encoding (CITE). Roy and Bhattacharjee proposed a Local Gradient Fuzzy Pattern (LGFP) for sketch to photo matching. This was then followed by Peng *et al.* [45], [46] who proposed a method that considers the facial spatial structure while extracting the features for matching.

Most of the proposed methods in inter-modality approach describe the image using local features concatenation as in [15]. First, the image is divided equally into patches. Then, the feature extraction is done on each patch (i.e., local features). These local features are then concatenated to build a full feature vector that represents the image. Recently, Haghighat *et al.* [25] fuse two feature vectors to obtain a single feature vector. This is to make the feature vector more discriminative than the original separated feature. The proposed method uses a popular local descriptors called Histograms of Oriented Gradients (HOG) [47] and global descriptors called Gabor Wavelet (GW) features [48]. This GW features is said to be robust against illumination change and hence make it relevant to solve one of the aforementioned problems. They fused the features based on Canonical Correlation Analysis (CCA) technique proposed by [20]. Our proposed method (illustrated in Fig. 1) is based on this approach. We cascade another matching block (after the first matching block) that is able to cater slight shape exaggeration effects. Eventually, the score from the first block and the second block are fused to give the final score. Patch of Interest (PoI) is also introduced here to deal with the processing time consumption, higher feature dimension and noise inclusion.

III. CANONICAL CORRELATION ANALYSIS FEATURE FUSION AND MATCHING

In order to make use of local descriptors, the face images need to be aligned. Directly match unaligned image pairs will result in poor matching accuracy. The face images are normally aligned by performing translation, rotation and scaling such that the two eyes are positioned at a fixed reference points [14], [15]. Klare *et al.* [15] reported that the outer regions (e.g., hairline and chin) of forensic sketches are more salient than the inner regions (e.g., eyes, nose, and mouth). It means that the outer region is more discriminative than the inner regions. Following to these findings, here, three fiducial points from the outer region are used to align the face images. The points are at left and right face edge (i.e., on horizontal line of the eyes) and chin tip. The alignment is done by employing affine transformation based on these three points.

Motivated by the recent findings, Haghighat *et al.* [25] managed to produce a more discriminative feature vector after combining two different input features. They fused the features based on Canonical Correlation Analysis (CCA) technique proposed in [20]. The proposed method uses a popular local descriptors called Histograms of Oriented Gradients (HOG) [47] and global descriptors called Gabor Wavelet (GW) features [48]. This GW features is robust against image that is effected by lighting variation. Hence, make it suitable for this application. This paper extends the approach (later becomes the first matching block) such that it can cater slight shape exaggeration effects. The details of HOG feature extraction, GW feature extraction and CCA feature fusion technique are provided in the following subsections.

A. HISTOGRAM OF ORIENTED GRADIENT

The first feature for the feature fusion is Histogram of Oriented Gradient (HOG) that is introduced by Dalal and Trigg [47]. This descriptor is initially proposed for human detection and has been tested to work well in face

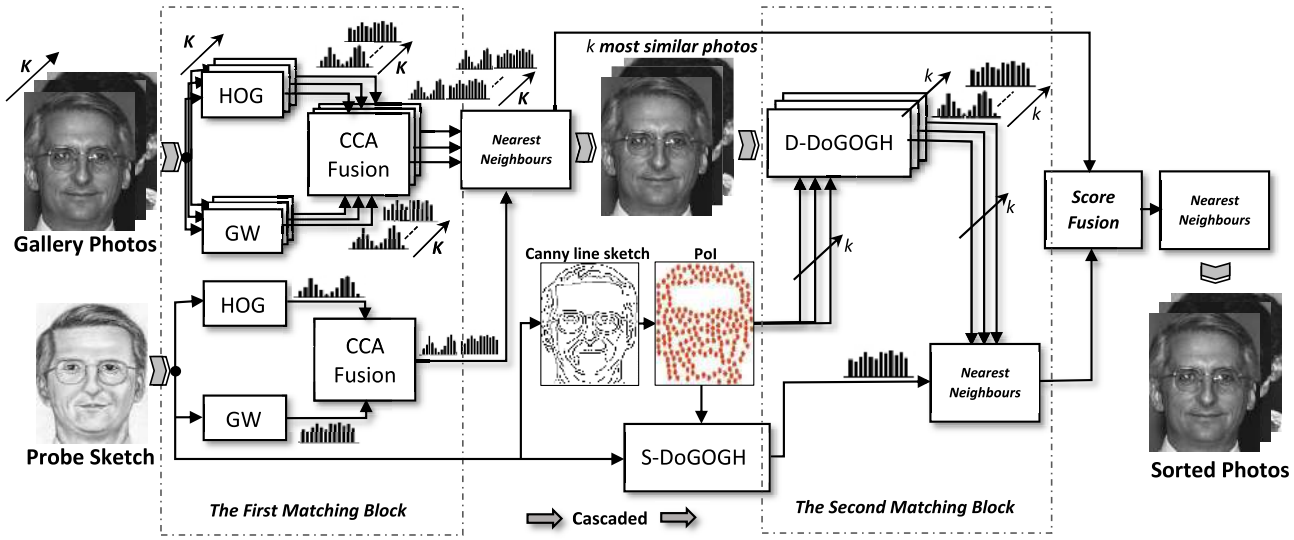


FIGURE 1. The proposed face sketch image retrieval system. Histogram of Oriented Gradient (HOG) and Gabor Wavelet (GW) features are fused by Canonical Correlation Analysis (CCA) in the first matching block. Then, the Static Difference of Gaussian Oriented Gradient Histogram (S-DoGOGH) and Dynamic Difference of Gaussian Oriented Gradient Histogram (D-DoGOGH) features are extracted on the Patch of Interest (PoI) in the second matching block. The scores from both matching blocks are fused and the short listed photos from the first matching block are re-sorted as the final result. *K* is the size of gallery and *k* is the number of short listed photos.

recognition [49]. To extract the feature, let $I(x, y)$ be the grayscale aligned image (i.e., using three fiducial points). Next, the histogram of the gradient vectors is computed on each pixel of the image. The following equation formulates the gradient vectors:

$$\nabla I(x,y) = \begin{bmatrix} G_x(x,y) \\ G_y(x,y) \end{bmatrix} = \begin{bmatrix} \frac{\partial I(x,y)}{\partial x} \\ \frac{\partial I(x,y)}{\partial y} \end{bmatrix} \tag{1}$$

This gradient vectors $\nabla I(x, y)$ are used to further compute the gradient orientation $\theta I(x, y)$ and magnitude $|\nabla I(x, y)|$ of each pixel in the image using the following formula:

$$\theta I(x,y) = tan^{-1}\frac{G_y(x,y)}{G_x(x,y)} \quad \in [-\pi, \pi] \tag{2}$$

$$|\nabla I(x,y)| = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{3}$$

Let $F^H = [f_a, \ldots, f_M]$ be the extracted feature vector of image $I(x, y)$ where $a = 1, 2 \ldots, M$. To extract the feature locally, the gradient orientation $\theta I(x, y)$ and magnitude $|\nabla I(x, y)|$ are divided equally into a set of M small overlapping patches. Each patch contains 2×2 cells. On each cell, the gradient magnitude $|\nabla I(x, y)|$ of all pixel positions within the cell is accumulated in evenly spaced bins ranges from -π to π (i.e., to cater light-to-dark and dark-to-light transitions). The accumulation is based on it respective orientation $\theta I(x, y)$. Each cell produces an oriented gradient histogram, c to build $f_a = [c_{11}, c_{12}, c_{21}, c_{22}]$. Then, the feature vector f_a is normalized using L_1-norm normalization scheme as follows:

$$f'_a = \frac{f_a}{\|f_a\|_1 + \epsilon} \tag{4}$$

where ϵ is a small constant. The extraction process is repeated for all patches across the image and these feature vectors are concatenated to build a HOG descriptor. Algorithm 1 shows the extraction details.

Algorithm 1 HOG Feature Extraction Method

Input: Aligned face image $I(x, y)$.
***Step* 1**: *Preprocessing*. Convert the image into grayscale.
***Step* 2**: *Orientation and Magnitude Computation*. Compute orientation $\theta I(x, y)$ and magnitude $|\nabla I(x, y)|$ of each pixel on the $I(x, y)$ image using Equation 2 and 3, respectively.
***Step* 3**: *Extract Features*. Divide the orientation $\theta I(x, y)$ and magnitude $|\nabla I(x, y)|$ 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 $f_a^h = []$.
2: $f_a^h \leftarrow |\nabla I_a|$ according to θI_a.
3: Normalize the f_a^h using L_1-norm as in Equation 4.

Concatenate these feature vectors $f_a^{\prime h}$ to build a HOG descriptor $F^H = [f_a^{\prime h}, \ldots, f_M^{\prime h}]^T$.
Output: F^H.

B. GABOR WAVELET FEATURES

The second feature for the feature fusion is Gabor Wavelet (GW) features. This feature is robust against illumination change and has been proven to work effectively in face recognition [48], [50]. Moreover, it also appropriate for texture representation and segmentation [51], [52].

The complex Gabor function in spatial domain is given by

$$B(x, y) = S(x, y)W(x, y) \tag{5}$$

where $S(x, y)$ is a complex sinusoid (i.e., the carrier) and $W(x, y)$ is a 2D Gaussian function (i.e., the envelope). The carrier $S(x, y)$ and the envelope $W(x, y)$ are defined as

$$S(x, y) = exp(j2\pi fx' + \phi) \tag{6}$$
$$W(x, y) = L\ exp(-(\alpha^2 x'^2 + \beta^2 y'^2)) \tag{7}$$

and

$$x' = x\ \cos\ \theta + y\ \sin\ \theta$$
$$y' = -x\ \sin\ \theta + y\ \cos\ \theta \tag{8}$$

where f is the central frequency of the carrier, ϕ is the phase of the carrier, L is the scaling factor for the magnitude of the Gaussian envelope and θ is the rotation angle of the Gaussian envelope. α and β are the sharpness of the Gaussian major axis and minor axis, respectively. The details are explained in [50].

To extract the feature, let $I(x, y)$ be the grayscale aligned image (i.e., using three fiducial points). The feature is obtained by convolving the image $I(x, y)$ with the Gabor kernel $B(x, y)$ as defined in (5). The convolution process is defined as

$$I'_{u,v}(x, y) = I(x, y) * B_{u,v}(x, y) \tag{9}$$

where $*$ denotes the convolution operator and $I'_{u,v}(x, y)$ denotes the convolved images at different orientations u and scales v. The magnitude $|I'_{u,v}(x, y)|$ of each $I'_{u,v}(x, y)$ is then downsampled, normalized and reshaped (by concatenating the rows or columns [53]) to be the feature vectors $f^b_{u,v}$. Eventually, the Gabor Wavelet (GW) feature, F^B is built by concatenating the feature vectors $f^b_{u,v}$ for all orientations and scales. This is referred to as a global feature extraction method because there is no patch by patch feature extraction. Algorithm 2 shows the feature extraction method in step by step manner.

Algorithm 2 GW Feature Extraction Method

Input: Aligned face image $I(x, y)$.
***Step* 1**: *Preprocessing*. Convert the image into grayscale.
***Step* 2**: *Wavelet Generation*. Create Gabor Wavelet kernels $B_{u,v}(x, y)$ by using Equation 5 as many as $U \times V$ (number of orientation $\times$ number of scale). Let $u = 1, \ldots, U$ and $v = 1, \ldots, V$.
***Step* 3**: *Convolution Process*. Convolve the image $I(x, y)$ with each Gabor Wavelet kernel in $B_{u,v}(x, y)$ as in Equation 9. Resulting $I'_{u,v}(x, y)$.
***Step* 4**: *Build Feature Vectors*. Compute the magnitudes $|I'_{u,v}(x, y)|$ of $I'_{u,v}(x, y)$. Then, downsample and normalize it to zero mean and unit variance. Reshape the result by column-wise concatenation to build the feature vectors $f^b_{u,v}$.
***Step* 5**: *Concatenation Process*. Concatenate these feature vectors, $f^b_{u,v}$ for all orientations and scales to build Gabor Wavelet (GW) feature vector $F^B = [(f^b_{1,1})^T, \ldots, (f^b_{U,V})^T]^T$.
Output: F^B.

C. CANONICAL CORRELATION ANALYSIS FEATURE FUSION

To make the feature vector more discriminative, two feature vectors are combined into a single feature vector as proposed by Haghighat *et al.* [25]. They fused the features based on Canonical Correlation Analysis (CCA) technique proposed by [20]. The two vectors are HOG, F^H and GW, F^B. These vectors come with high dimension. Directly computing CCA on high dimensional feature vector is computational expensive. Principle Component Analysis (PCA) [54] is an effective dimensionality reduction technique used to reduce the feature vector dimension. Hence, it is employed here. The correlation is then analyzed using CCA.

In multivariate analysis, CCA is used to maximize the correlation between two sets of variables. It identifies and quantifies the correlation between the sets based on a linear combination of the variables. Here, suppose that $X \in \mathbb{R}^{p\times n}$ and $Y \in \mathbb{R}^{q\times n}$ are two matrices, each contains n training feature vectors from two different descriptors. Let $X = [F^H_1, \ldots, F^H_n]$ be the feature matrix for HOG and $Y = [F^B_1, \ldots, F^B_n]$ be the feature matrix for GW. Based on these two matrices, the covariance matrix is defines as

$$\Sigma = \begin{bmatrix} cov(X, X) & cov(X, Y) \\ cov(Y, X) & cov(Y, Y) \end{bmatrix} = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \tag{10}$$

where Σ_{XX} and Σ_{YY} denote the within class covariance matrices of X and Y, respectively. Σ_{XY} and Σ_{YX} denote the between class covariance matrices of X and Y, respectively. The CCA attempts to find the coefficient vectors w_X and w_Y for the linear combinations of X and Y (i.e., $X' = w_X^T X$ and $Y' = w_Y^T Y$) such that it maximizes the correlation between the two. For that, the coefficient vectors w_X and w_Y need to be computed such that the correlation (11) between X' and Y' is maximized.

$$Corr(X', Y') = \frac{w_X^T \Sigma_{XY} w_Y}{\sqrt{w_X^T \Sigma_{XX} w_X}\sqrt{w_Y^T \Sigma_{YY} w_Y}} \tag{11}$$

The coefficient vectors w_X and w_Y (i.e., the transformation matrices) are the eigenvectors with the maximum eigenvalue of the matrix $\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$ and $\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$, respectively. The transformed features (i.e., X' and Y') are then fused by concatenation as in (12) into a single feature vector, F^{HB}.

$$F^{HB} = [X', Y'] = [w_X^T X, w_Y^T Y] \tag{12}$$

D. THE FIRST MATCHING BLOCK

In order to match the feature vector of a sketch with the feature vectors of a set of photos, let $X^P = [F^{H(P)}_1, \ldots, F^{H(P)}_n]$ and $Y^P = [F^{B(P)}_1, \ldots, F^{B(P)}_n]$ be the extracted feature vectors from the photos using HOG and GW, respectively. From (12), the feature vectors are fused to become $F^{HB(P)} = [w_X^T X^P, w_Y^T Y^P]$. Now, let $X^S = F^{H(S)}$ and $Y^S = F^{B(S)}$ be the extracted feature for the sketch using HOG and GW,

respectively. From (12), by using the derived transformation matrices w_X and w_Y (i.e., based on X^P and Y^P), the sketch features are transformed and fused to become $F^{HB(S)} = [w_X^T X^S, w_Y^T Y^S]$. Then, the L_2-distance is measured between the sketch feature vector $F^{HB(S)}$ and the photo feature vectors $F^{HB(P)}$. The classification here is based on the nearest neighbour (i.e., based on the smallest distance). Refer to Algorithm 3.

Algorithm 3 First Block Matching Algorithm

Input: The fused sketch feature vector $F^{HB(S)}$; The fused photo feature vector $F_g^{HB(P)}$ where $g = 1, 2, \ldots, K$. K is the size of gallery.

***Step* 1**: Calculate the L_2-distance d'_g between $F^{HB(S)}$ and $F_g^{HB(P)}$ as follows:

$$d'_g = \|F^{HB(S)} - F_g^{HB(P)}\|_2^2 \tag{13}$$

***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 distances, d'_{g_s} and indexes, g_s.

After the indexes have been sorted, nearest neighbour is used to short list k number of photos based on the first k sorted indexes g_s. The selected indexes are now pointing to the photos that have the highest similarity against the probe sketch. Later, the second matching block performs the match based on these photos (a total of k) for rank-1 accuracy improvement.

IV. PATCH OF INTEREST DYNAMIC LOCAL FEATURE MATCHING

In the intra-modality approach, a pseudo-photo or pseudo-sketch (i.e., pseudo-image) is synthesized using a synthesizing algorithm. The algorithm is normally complex. It synthesize the probe image such that the generated pseudo-image looks visually similar to that the probe image (including the shape position) but in different modality. This is intuitively naive. Matching this kind of pseudo-image with its corresponding pair using insensitive shape exaggeration descriptors may result in a low matching rate even though both images are in the same modality. It even worse when the face sketch is rendered with higher degrees of shape exaggeration. Similarly, for inter-modality approach, shape exaggeration effects are usually ignored. In this approach, local feature extraction strategy is among the popular method to extract the features [15]. The features are extracted from static patches independently for the sketch and for the photo. This may cause improper feature representation for the current patch when the corresponding sketch contains shape exaggeration. Thus, reducing the retrieval accuracy.

In order to ensure a better retrieval rate, there are two observed problems that require proper treatment: illumination difference and shape exaggeration. To tackle the illumination problem, we propose so that the image is represented by Difference of Gaussian Oriented Gradient Histogram (DoGOGH) feature [26]. This is because DoGOGH has been proven to perform effectively on matching face sketch to photo with illumination effects [26]. For the latter problem, we propose such that the local feature vector is extracted dynamically (i.e., based on a reference patch) using dynamic local feature extraction method [44]. This is to cater a possibility of the exaggerated shape resides in the neighbouring patches. Hence, matching is made up on the appropriate pairs. The fact that forensic sketch has some degrees of abstraction with less accurate details and has non-linear mapping properties, directly match the entire extracted features from local patches seems to be unfair. It may consume longer matching time due to higher feature dimension and may also match unnecessary patch pairs (i.e., noise inclusion). To address this, we propose Patch of Interest (PoI) dynamic local feature extraction and matching method.

A. PATCH OF INTEREST

Visualizing a face of unknown person in mind without seeing it requires a full face descriptions that are emphasized on the unique facial features of that person. Similarly, generating a forensic sketch also requires all those features to be elicited from the eyewitness. Based on the observation, forensic sketch (i.e., hand-drawn) is literally a composition of strokes (with a certain curve, intensity and thickness to make up a shape) and the blurring shadows (mostly stressed around the shapes to realize the sketch). Matching only at Patch of Interest (PoI) may be a solution to improve the processing time, reduce feature dimension and minimize noise inclusion.

In this work, PoI is defined as a local patch with the centre point positioned at any point on the edges of intensity face image. In order to identify the point, an edge detector is required. Based on the study and comparison of popular edge detection method done by [55], the article reported that Canny edge detection algorithm performs better than others (i.e., Laplacian of Gaussian, Sobel, Prewitt and Robert). Based on this result, the Canny edge detection algorithm is extended in this work to output the centre points of the PoI. Hence, be another proposed method. The algorithm is outlined in Algorithm 4. In brief, a grey sketch image (as in Fig. 2 (a)) is loaded and transformed into a line sketch image (as in Fig. 2 (b)) using Canny edge detection (i.e., step 1 to 4). Then, the PoI centre points are localized using the extended steps (i.e., step 5 to 9).

B. DYNAMIC LOCAL FEATURE EXTRACTION ON PoI

Generated facial sketch is usually exposed to some degree of shape exaggeration (especially on forensic sketch). This may cause the local feature to experience the same exaggeration effects, thus lower the retrieval rate. If the feature is extracted from a proper local position, the retrieval rate could be increased due to the fact that the feature matching is made on the true pairs. To address this, here, we use dynamic local feature extraction as described in [44]. In the context of matching sketch to photo, dynamic local feature

Algorithm 4 PoI Center Points Localization

Initialize δ = selected distance, $\iota = 0$, T_l and T_h. T_l and T_h are the lower and higher threshold, respectively. λ is defined as the length of vector d.

***Step* 1**: *Image Smoothing*. Convolve the image $I(x, y)$ with a 5×5 Gaussian filter $G_\sigma(x, y)$ to remove the noise.

$$G_\sigma(x, y) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{x^2+y^2}{2\sigma^2}} \tag{14}$$

$$\hat{I}(x, y) = I(x, y) * G_\sigma(x, y) \tag{15}$$

***Step* 2**: *Orientation and Magnitude Computation*. Convolve the smoothened image $\hat{I}(x, y)$ with a Sobel kernel in horizontal and vertical directions to obtain first derivative in horizontal direction $G_x(x, y)$ and vertical direction $G_y(x, y)$, respectively. Then, compute the direction and magnitude for each pixel as follows:

$$\theta G(x, y) = tan^{-1} \frac{G_y(x, y)}{G_x(x, y)} \quad \in [-\pi, \pi] \tag{16}$$

$$|G(x, y)| = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{17}$$

***Step* 3**: *Non-Maximum Suppression*. In the magnitude $|G(x, y)|$, at each pixel, remove (i.e., set to zero) the pixel if it is not a local maximum among its neighborhood in the direction of gradient $\theta G(x, y)$. Let the result be $\hat{G}(x, y)$.

***Step* 4**: *Hysteresis Thresholding*. In $\hat{G}(x, y)$, pixel with magnitude less than T_l is discarded (i.e., set to zero) and pixel with magnitude in between T_l and T_h is discarded (i.e., set to zero) if it is not connected to the pixel with magnitude higher than T_h. All other pixels is set to one. The result will be a binary image, $I_{BW}(x, y)$.

***Step* 5**: *Points Initialization*. All the white pixel coordinates from $I_{BW}(x, y)$ are initialized as $\zeta = [\boldsymbol{x}, \boldsymbol{y}]$ where $\boldsymbol{x} = \left[x_1, x_2, \ldots, x_j, \ldots, x_\rho\right]^T$, $\boldsymbol{y} = \left[y_1, y_2, \ldots, y_j, \ldots, y_\rho\right]^T$ and $j = 1, 2, \ldots, \rho$. ρ is the vector length.

***Step* 6**: *Point Selection*. Extract out the first point from ζ into D, where $D_{\iota+1} = [x_1, y_1]$ (*as the first center of Patch of Interest*) and reconstruct $\zeta = [\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}]$ where $\hat{\boldsymbol{x}} = \left[x_2, \ldots, x_\rho\right]^T$ and $\hat{\boldsymbol{y}} = \left[y_2, \ldots, y_\rho\right]^T$. Hence, $\zeta = \left[[\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_i, \ldots, \hat{x}_\lambda]^T, [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_i, \ldots, \hat{y}_\lambda]^T\right]$ where $i = 1, 2, \ldots, \lambda$ and $\lambda = \rho - 1$.

***Step* 7**: *Distance Measure*. Compute the distances, d_i between the selected point (in Step 6) and all other points.

$$d_i = \sqrt{(x_1 - \hat{x}_i)^2 + (y_1 - \hat{y}_i)^2} \tag{18}$$

***Step* 8**: *Close Points Removal*. Update ζ by removing all points that have distance lesser than δ.

$$\zeta = [x_j, y_j] = [x_i, y_i] = \begin{cases} [\hat{x}_i, \hat{y}_i], & \text{if } d_i \geq \delta \\ remove, & \text{otherwise} \end{cases} \tag{19}$$

Algorithm 4 *(Continued.)* PoI Center Points Localization

where $j = 1, 2, \ldots, \rho_{new}$ is the new index of i due to the re-indexing process. $\rho \leq \lambda$.

***Step* 9**: Repeat from Step 6 while ζ is not empty.

Output: $D = [D_1, D_2, \ldots, D_M]^T$ where M is the number of valid points (i.e., later be the number of patches) extracted from ζ.

FIGURE 2. Example of the interest points selection; (a) the input sketch, (b) line sketch, and (c) the plotted points after selection.

extraction extracts feature vector from local patch (of the photo) dynamically based on a reference feature vector (extracted from sketch) at the same local position.

In order to extract the feature, we utilize DoGOGH descriptor [26]. Let $F^S = [f_m^S, \ldots, f_M^S]$ and $F^P = [f_m^P, \ldots, f_M^P]$ denote the feature vector extracted from the sketch $I^S(x, y)$ and photo $I^P(x, y)$, respectively. Here, $m = 1, \ldots, M$ where M is the total number of Patch of Interest (PoI). For the sketch features, the static local feature extraction method is employed (i.e., S-DoGOGH) as in [26]. We call this method as static extraction method because the features are extracted out from patches at fixed local position. In addition to that, another local feature extraction method as in [44] is employed (i.e., D-DoGOGH) to extract dynamic feature from the photo. To dynamically extract the feature vector f_m^P from photo, let consider $F'^P = [f_{ml}^P, \ldots, f_{ML}^P]$. Here, $l = 1, \ldots, L$ where L is the total number of patch candidates (i.e., the target and its neighbouring patches). First, localize the PoI center points D by using Algorithm 4 on the sketch $I^S(x, y)$. These points (represent the center pixel of the patches) are assigned as the target points. Then, for the sketch, extract DoGOGH feature F^S on these points (which later be the reference feature for dynamic extraction on photo). Similarly, for the photo, extract DoGOGH feature F'^P on these points and its neighbouring points as what has been proposed in [44]. Based on these feature vectors, the distances between f_m^S and f_{ml}^P are computed using L_1-distance measure. The feature vector from f_{ml}^P that has the smallest distance against f_m^S is chosen to represent the current patch feature vector f_m^P. This process is repeated for all patches across the image to construct F^P.

C. THE SECOND MATCHING BLOCK

In order to improve the rank-1 matching accuracy, the k short listed images from the first matching block are re-matched

using Patch of Interest (PoI) dynamic local feature. Let F^P be the dynamic extracted feature vector for photo and F^S be the static extracted feature vector from the sketch, respectively. Now, let $F_g^P = [F_1^P, \ldots, F_k^P]$ denote the short listed photo features from the gallery. Next, the L_1-distance (i.e., the score) is measured between the sketch feature vector F^S and the photo feature vectors F_g^P. Fusion of the two scores (from the first matching block and second matching block) are done by first normalizing the scores using min-max normalization. In this case, the gallery size has been reduced to contain a total of k subjects. Therefore, there will be k scores that represent the distances between the probe sketch and all the photos in the gallery. Then, the sum-of-scores method is used to fuse the normalized scores from the first matching block and second matching block. Refer Algorithm 5 for the details. Eventually, the retrieval rate is computed based on the new score.

Algorithm 5 : Second Block Matching Algorithm

Input: Sketch feature vector F^S; Photo feature vector F_g^P, where $g = 1, 2, \ldots, k$ and k is the number of short listed photos; and k short listed score $d'_g = d'_{g_s}$.

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

$$d''_g = \|F^S - F_g^P\|_1 \tag{20}$$

***Step* 2**: Normalize the scores d'_g and d''_g using min-max normalization technique as follows:

$$\hat{d} = \frac{d - d_{min}}{d_{max} - d_{min}} \tag{21}$$

***Step* 3**: Fuse the min-max normalized scores $\hat{d}'_g$ and $\hat{d}''_g$ (from Step 2) using sum-score fusion technique as follows:

$$SF_g = \hat{d}'_g + \hat{d''}_g \tag{22}$$

Output: The fused score, SF_g.

V. EXPERIMENTS

In order to evaluate the effectiveness of the proposed method, a dataset with illumination and shape exaggeration effects should be considered. For that, two benchmark datasets were used; CUHK Face Sketch Database (CUFS) and CUHK Face Sketch FERET Database (CUFSF). The sketches in both datasets were observed to have been rendered with shape exaggeration. Between the two, the sketches in the CUFS dataset has almost no shape exaggeration and the sketches in the CUFSF dataset has some degrees of shape exaggeration as to mimic real forensic sketches. These datasets are in *Viewed Sketch* category in which the artist sketched the face while viewing the photo. Fig. 3 (a) to (d) show the example sketch and its corresponding photo for *Viewed Sketch*. The evaluation was eventually extended to *Semi-Forensic Sketch* (IIIT-Delhi Semi-forensic Sketch Database (IIIT-D)) and *Forensic Sketch* (PRIP Hand-Drawn Composite (PRIP-HDC) dataset (PRIP)) datasets. This is to investigate the feasibility of the proposed method beyond the *Viewed Sketch*. The *Semi-Forensic Sketch* is the sketch drawn by an artist without the descriptions from others but solely from their memory after viewing the photo while the *Forensic Sketch* is rendered based on the verbal descriptions given by the eyewitness. Fig. 3 (e) and Fig. 3 (f) show the example sketch and its corresponding photo for *Semi-Forensic Sketch* and *Forensic Sketch*, respectively. The fact that the proposed method does not require a training phase, thus all available samples were used for testing.

CUHK Face Sketch Database (CUFS) was prepared by [3], [7]. It contains 606 image pairs from CUHK student dataset [56] (188 image pairs), AR dataset [57] (123 image pairs) and XM2VTS dataset [58] (295 image pairs). The photos are in frontal pose with a neutral expression, and under normal lighting condition. Another database is CUHK Face Sketch FERET Database (CUFSF) [7], [29]. It is drawn based on 1,194 photos from the FERET database [59]. The sketches are generated with shape exaggeration and some of the photo are exposed to lighting variation. IIIT-Delhi Semi-forensic Sketch Database (IIIT-D) database was developed by [60]. It comprises 140 Semi-Forensic Sketch pairs; 65 from IIIT-D student and staff, 34 from FG-NET and 41 from CUHK database. Out of all, only 106 (IIIT-D student and staff + CUHK) semi-forensic image pairs are available for testing in this work. The Pattern Recognition and Image Processing (PRIP) Hand-Drawn Composite (PRIP-HDC) dataset is another database used in the evaluation. It was prepared by [61]. It consists of 265 hand-drawn Forensic Sketch pairs. Of the 265 total sketches, 116 were drawn by [1] and [2], 102 sketches were provided by Pinellas County Sheriff's Office (PCSO) and Michigan State Police. The remaining 47 sketches were downloaded from the Internet. Due to the intellectual property (IP) issue, only 47 sketches pair from the Internet is released to researchers.

A. EXPERIMENTAL SETUP

The sketches and photos are affine transformed to a fixed reference points $\vec{tr} = [r_1, r_2, r_3] = [(15, 80), (126, 80), (71, 161)]$ based on the defined fiducial points. Then, the images are cropped to size 175×140 pixels. In the first matching block, fifty percent overlapping patch of size 16×16 is used for HOG feature extraction in this experiment. With this setting, a total of 320 patches (i.e., $M = 320$) per image is obtained. For HOG computation, the number of orientation bins are set to 18 and each patch is divided into 4 cells to yield a $72M$ concatenated feature descriptor. Another descriptor is built based on Gabor Wavelet (GW) feature. Here, forty Gabor filters are used (i.e., five scales and eight orientations) and the feature images are downsampled by a factor of four. In the second matching block, for DoGOGH computation, the two different widths used in the Gaussian kernel $G_\sigma(x, y)$ are $\sigma_1 = 1$ and $\sigma_2 = 2$ [26]. The neighbouring patch searching coverage d_p in [44] is set to 4 pixels while the selected distance δ in Algorithm 4

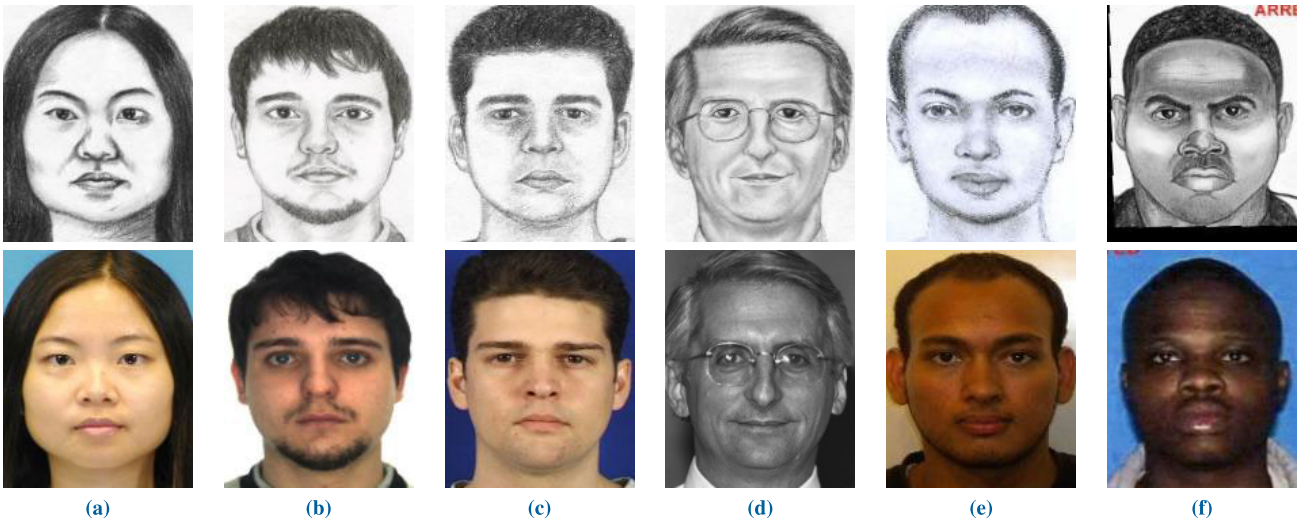


FIGURE 3. Examples of viewed, semi-forensic, and forensic sketch pair. Viewed sketch: (a) CUHK, (b) AR, (c) XM2VTS, (d) FERET. Semi-Forensic sketch: (e) IIIT-D. Forensic sketch: (f) PRIP. Note that (a), (b) and (c) are from CUFS dataset while (d) is from CUFSF dataset.

is set to 8 pixels. The values of hysteresis thresholding in Algorithm 4 are chosen by using a histogram approach (i.e., MATLAB default settings), where the gradient magnitude is binned into 64 levels, and the threshold is based on a percentile of the gradient values (the 70% level is the high threshold), and 28% (0.4 of the 70%) is the level for the lower threshold. The number of short listed photos k in Algorithm 5 is assigned to 10. All other settings and parameters used in this 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. RESULTS

Here, a Cumulative Match Curve (CMC) (as in Algorithm 6) is used to evaluate the performance. This is a commonly used evaluation tool by most of the researchers in this field [11], [14], [15], [18], [61]–[67]. It measures the percentage of correct identity cumulatively across the ranks. Rank-1 accuracy simply means the percentage of correct match that is merely based on the smallest distance (i.e., similar to retrieval rate). Rank-k accuracy gives the retrieval rate that the correct match can be retrieved within the first k smallest distances. For example, if rank-k percentage is at 100%, it simply means that the proposed method is able to short list k face candidates without mistake (i.e., the correct match resides in the short listed photos).

In order to show the retrieval rate improvement of the proposed method progressively from the baseline descriptor (i.e., Histogram of Oriented Gradient (HOG)), the performance comparison is generated in Fig. 4. Fig. 4a and Fig. 4b show the retrieval rate (based on CMC) for the first ten ranks obtained on CUFS and CUFSF datasets, respectively. From the Fig. 4a, the retrieval rate improvement pattern is less obvious because the dataset has very less illumination effects and shape exaggeration. But, the proposed method shows a comparable accuracy on that dataset. On the other dataset, Fig. 4b indicates progressive retrieval rate improvement pattern from Histogram of Oriented Gradient (HOG) descriptor towards the proposed method (i.e., CCA Fusion + D-DoGOGH on PoI). Note that the CUFSF dataset is

Algorithm 6 : Cumulative Match Curve (CMC) Algorithm

Input: Distance Matrix, $Dist^{S\times P}$ where S is the number of rows (sketches) and P is the number of columns (photos). The ranks $Rank_j^{1\times P} = [Rank_1, \ldots, Rank_P] = [0, \ldots, 0]$ (i.e., Rank-1 to Rank-P is set to zero). Initialize $s = 1$.

***Step* 1**: *Distance Labelling*. Let $dist_s^{1\times P}$ be the s^{th} row of $Dist^{S\times P}$ and $lbl_s^{1\times P}$ be the labelling indexes ranges from 1 (smallest distance) to P (largest distance). Label the $lbl_s^{1\times P}$ according to the distances in $dist_s^{1\times P}$.

***Step* 2**: *Score Binning*. Let all the ranks $Rank_j^{1\times P}$ be the bins and let $lbl_s^{1\times P} = l_{s,p} = [l_{s,1}, \ldots l_{s,P}]$. Get the index value i_v from $lbl_{s,p}$ at $s = p$ (i.e., the correct pair). Add 1 to the $Rank_j$ (j^{th} bin) where $j = i_v$ as the following:

$$Rank_j = Rank_j + 1 \tag{23}$$

***Step* 3**: Repeat Step 1 and Step 2 for the next s until $s = S$.

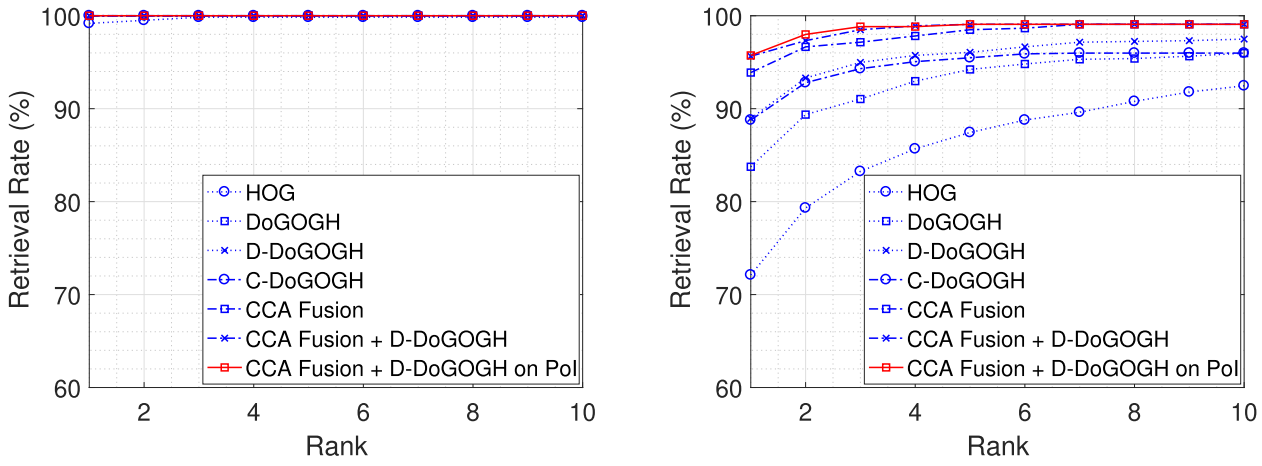
***Step* 4**: *Score Accumulation*. Accumulate the score by using the following formula iteratively from $j = 1$ to P

$$Rank_j = Rank_{j-1} + Rank_j, \quad Rank_0 = 0 \tag{24}$$

***Step* 5**: *Percentage Computation*. Compute the ranks percentage by using the following formula:

$$Rank_j = \frac{Rank_j}{S} \times 100, \quad 0 < j \leq P \tag{25}$$

Output: The ranks $Rank_j^{1\times P}$.



(a) CUFS dataset. **(b)** CUFSF dataset.

FIGURE 4. Retrieval rate (CMC) comparison from the Histogram of Oriented Gradient (HOG) [47] towards the proposed method evaluated on (a) CUFS dataset and (b) CUFSF dataset. HOG is extended to Difference of Gaussian Oriented Gradient Histogram (DoGOGH) [26], then DoGOGH is extended further to Cascaded DoGOGH (C-DoGOGH) [44]. HOG and Gabor Wavelet (GW) features are fused using CCA (CCA Fusion) and finally is the proposed method (CCA Fusion + D-DoGOGH on PoI).

prepared with shape exaggeration for the sketches and with illumination effects for the photos. By using the proposed method, the rank-1 accuracy is improved to 95.48% as a prove that the proposed method can cater for illumination effects and shape exaggeration.

Next, the accuracy comparison of the proposed method with the state-of-the-art method is tabulated in Table 1. From the table, it can be clearly seen that the proposed method outperforms the other methods (i.e., based on the reported accuracy from the respective publications). Note that the proposed method does not require any training or synthesizing process and thus suitable for real time application. The proposed method belongs to inter-modality approach. Due to that fact, the comparison is also made on several inter-modality approaches that include Multi-feature Cannonical Correlation Analysis (MCCA), Coupled Information-Theoric Encoding (CITE), Scale-Invariant Feature Transform (SIFT) + Multiscale Local Binary Pattern (MLBP), Local Radon Binary Pattern (LRBP), and Local Difference of Gaussian Binary Pattern (LDoGBP). The fact that deep learned network can also extract effective feature, we adopted the Siamese CNN architecture based on the pre-trained models (i.e., two identical networks) for feature extraction and we compute the similarity of the features using L_1-distance as what has been evaluated by [44]. The accuracies were inserted in Table 1 for additional comparison.

From the rank-1 accuracies summarized in Table 1, CCA Fusion with D-DoGoGH on PoI matching method has outperformed the state-of-the-art methods when tested on viewed sketch datasets (i.e., CUFS and CUFSF). However, for real world application, forensic sketch is used instead of the viewed sketch. In order to test the feasibility of the proposed method beyond the viewed sketch, the experimental evaluation was extended to semi-forensic dataset (i.e., IIIT-D) and the forensic dataset (i.e., PRIP). Table 2 tabulates the results. The comparison was made between CCA Fusion (our baseline) and CCA Fusion with D-DoGoGH on PoI matching methods. This is because the CCA Fusion matching method has outperformed the state-of-the-art methods (e.g., MCCA [68], CITE [29], LRBP [17], LDoGBP [69], and C-DoGOGH [44]) and thus be the baseline for the improved matching method (i.e., CCA Fusion with D-DoGoGH on PoI). From Table 2, the results demonstrate that the accuracy improvement for viewed sketch is not significant in comparison with the semi forensic and forensic sketch. The accuracy improvement of 66.77% is observed for forensic sketch to reach 10.64% rank-1 accuracy from 6.38%. Based on that observation, the feasibility of the proposed method on real forensic application can be considered as applicable for rank-1 accuracy improvement. To ensure results reproducibility, only retrievable 106 image pairs from the IIIT-D dataset and 47 image pairs from PRIP are considered. Based on this fact, in this evaluation, there is no comparison made to any state-of-the-art methods (i.e., for semi-forensic and forensic) as the testing samples used are mostly not retrievable (some are confidential as it belongs to the law enforcement authority), thus making up unfair comparison if the reported accuracies are included here.

To further evaluate the effectiveness of the proposed method on forensic sketch, we extend the number of gallery images (in PRIP dataset) from 47 to 1953 mugshots. This is done by combining all available photo samples (606 from CUFS, 1194 from CUFSF, 106 from IIIT-D and 47 from PRIP) in this work. The extension is made due to the fact that we want to mimic the real world forensic searching process in which the size of the extended gallery represent the mugshot database of a law enforcement agency. Similarly,

TABLE 1. Rank-1 accuracy comparison between the state-of-the-art methods on CUFS and CUFSF datasets. These accuracies were based on what have been reported in the respective literature except for the deep learning approaches marked with asterisk.

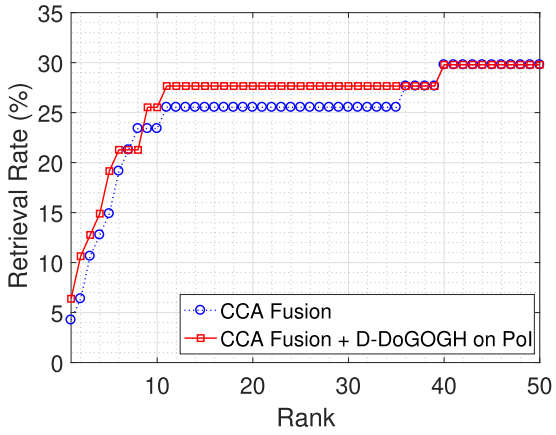
State-of-the-art methods	No. training samples	No. testing samples	Rank-1 accuracy (%)
CUHK Face Sketch Database (CUFS)			
Intra-Modality:			
MMRF + RS-LDA [7]	306	300	96.30
SNS-SRE [34]	306	300	96.50
TFSP + RS-LDA [10]	306	300	97.70
MrFSPS + RS-LDA [11]	306	300	97.70
S-FSPS + LDA [12]	306	300	99.10
Fully Convolutional Network [36]	88	100	100
Inter-Modality:			
SIFT + MLBP [15]	306	300	99.47
MCCA [68]	306	300	100
CITE [29]	306	300	99.87
LRBP [17]	0	606	99.51
LDoGBP [69]	0	606	96.53
*VGG Face Descriptor [40]	0	606	73.93
*Light CNN [41]	0	606	52.48
*ArcFace [42]	0	606	4.62
DoGOGH [26]	0	606	100
C-DoGOGH [44]	0	606	100
CCA Fusion + D-DoGOGH on PoI	0	606	**100**
CUHK Face Sketch FERET Database (CUFSF)			
Intra-Modality:			
TFSP + RS-LDA [10]	500	694	72.62
MrFSPS + RS-LDA [11]	500	694	75.36
S-FSPS + LDA [12]	500	694	72.19
Inter-Modality:			
MCCA [68]	300	894	92.17
CITE [29]	500	694	89.54
LRBP [17]	0	1194	91.12
LDoGBP [69]	0	1194	91.04
*VGG Face Descriptor [40]	0	1194	36.01
*Light CNN [41]	0	1194	38.27
*ArcFace [42]	0	1194	8.21
DoGOGH [26]	0	1194	83.75
C-DoGOGH [44]	0	1194	89.03
CCA Fusion + D-DoGOGH on PoI	0	1194	**95.48**

* Using similar evaluation method as in [44]

TABLE 2. Rank-1 accuracy comparison between CCA Fusion and CCA Fusion with D-DoGoGH on PoI matching methods evaluated on semi-forensic and forensic sketches.

Methods	Viewed (%)		Semi-Forensic (%)	Forensic (%)
	CUFS	CUFSF	IIIT-D	PRIP
CCA Fusion	100	93.89	33.02	6.38
CCA Fusion + D-DoGoGH on PoI	100	95.48	**42.45**	**10.64**
Improvement	0	1.69	28.56	66.77

in this evaluation, the comparison was made between CCA Fusion and CCA Fusion with D-DoGoGH on PoI matching methods. Here, the value of k in Algorithm 5 is set to 50. This is because we want to improve the retrieval rate accuracy based on the 50 short listed suspects. Fig. 5a shows the retrieval rate comparison for the first fifty ranks between the two methods. With the gallery size extended, the rank-1 accuracy for CCA Fusion matching method dropped from 6.38% to 4.26%. However, the CCA Fusion with D-DoGoGH on PoI matching method improve the rank-1 accuracy back to 6.38% (three out of forty-seven forensic sketches were identified at rank-1). In addition to that, the proposed method could roughly identify thirteen suspects if the consideration is made on the first eleven ranks. Overall, at rank-1, the results demonstrate that the CCA Fusion with D-DoGoGH on PoI matching method performs approximately 4% better than the CCA Fusion matching method on PRIP dataset and approximately 2% better than the CCA Fusion matching method on extended mugshot for PRIP dataset. Examples of forensic sketch with its corresponding photo match cases in which the proposed method improved the performance can be found in Fig. 5b.



(a) Retrieval rate for the first 50 ranks.

Method	Rank
CCA Fusion	3
CCA Fusion + D-DoGOGH on POI	1

Method	Rank
CCA Fusion	3
CCA Fusion + D-DoGOGH on POI	2

(b) Examples of retrieval rate improvement.

FIGURE 5. Retrieval rate for the proposed method on forensic sketch (from PRIP) with extended number of mugshots. The number of mugshots were extended from 47 (merely from PRIP) to 1953 (combining all available mugshot samples in this work: 606 from CUFS, 1194 from CUFSF, 106 from IIIT-D and 47 from PRIP). This is to ensure result reproducibility.

TABLE 3. Rank-1 accuracy and several average measurements comparison between CCA Fusion, CCA Fusion with D-DoGOGH and CCA Fusion with D-DoGOGH on PoI matching methods. CUFSF (1194 image pairs) dataset is used here.

Type of Measurement	Methods		
	CCA Fusion	CCA Fusion + D-DoGOGH	CCA Fusion + D-DoGOGH on PoI
Rank-1 Accuracy (%)	93.89	95.23	95.48
Average No. of Patch	-	320	167
Average Feature Dimension	-	23040	12024
Average Matching Time (s)	0.3495	0.3495 + 0.8205	0.3495 + 0.4830

In terms of average matching time, Table 3 tabulates the time comparison between CCA Fusion, CCA Fusion with D-DoGOGH and CCA Fusion with D-DoGOGH on PoI matching methods. From the table, CCA Fusion matching method requires 0.3495 seconds to perform the matching. Next, the result indicates that a combination of CCA Fusion and D-DoGOGH in a cascaded fashion is capable of improving the rank-1 accuracy by 1.43% but requires additional time of 0.8205 seconds to perform the matching. This is due to its high average number of patch and long feature dimension. The proposed method (i.e., CCA Fusion with D-DoGOGH on PoI) demonstrates better rank-1 accuracy of 95.48% with lesser average number of patch, feature dimension and matching time as compared to the CCA Fusion with D-DoGOGH matching method that give rank-1 accuracy at 95.23%.

VI. CONCLUSION

In this paper, we proposed a new approach for face sketch image retrieval using two stage matching blocks that are arranged in a cascaded fashion. Experiments on the CUFS and CUFSF databases have been conducted to evaluate the efficacy of the proposed method and the results indicate that the proposed method outperforms the state-of-the-art methods. To narrow down the most similar candidates from a big gallery images, in the first matching block, CCA is employed to fuse HOG and GW features into a single vector such that it becomes more discriminative (i.e., correlation between the features are maximized). Then, matching algorithm is executed based on these features. To improve the accuracy further, in the second matching block, we re-matched the sketch and photos using dynamically extracted local feature on its Patch of Interest (PoI). By doing this, the shape exaggeration is catered and the large feature dimension is reduced. In terms of the illumination effects, it is tackled by Gabor Wavelet (GW) feature in the first stage and Difference of Gaussian Oriented Gradient Histogram (DoGOGH) feature in the second stage. Similar to some other state-of-the-art methods, this proposed approach requires no training phase and thus computationally inexpensive. Next, the feasibility study of the proposed method on semi-forensic and real forensic datasets have been conducted and the results demonstrate that the proposed method is capable of improving the retrieval rate accuracy especially at rank-1.

REFERENCES

[1] L. Gibson, *Forensic Art Essentials: A Manual for Law Enforcement Artists*. Amsterdam, The Netherlands: Elsevier, 2008.

[2] K. T. Taylor, *Forensic Art Illustration*. Boca Raton, FL, USA: CRC Press, 2001.

[3] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Jun. 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. Gao, J. Zhong, J. Li, and C. Tian, "Face sketch synthesis algorithm based on E-HMM and selective ensemble," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 4, pp. 487–496, Apr. 2008.
[7] 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.
[8] W. Zhang, X. Wang, and X. Tang, "Lighting and pose robust face sketch synthesis," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 6316, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer, 2010.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] A. Radman and S. A. Suandi, "A superpixel-wise approach for face sketch synthesis," *IEEE Access*, vol. 7, pp. 108838–108849, 2019.
[14] B. Klare and A. K. Jain, "Sketch to photo matching: A feature-based approach," in *Proc. SPIE Conf. Biometric Technol. Hum. Identificat. VII*, 2010, Art. no. 766702.
[15] B. 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.
[16] H. K. Galoogahi and T. Sim, "Inter-modality face sketch recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 224–229.
[17] H. K. Galoogahi and T. Sim, "Face sketch recognition by local radon binary pattern: LRBP," in *Proc. 19th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2012, pp. 1837–1840.
[18] 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.
[19] M. A. A. Silva and G. C. Chavez, "Face sketch recognition from local features," in *Proc. 27th SIBGRAPI Conf. Graph., Patterns Images*, Aug. 2014, pp. 57–64.
[20] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognit.*, vol. 38, no. 12, pp. 2437–2448, Dec. 2005.
[21] N. Correa, T. Adali, Y.-O. Li, and V. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 39–50, Jun. 2010.
[22] J. Yang and X. Zhang, "Feature-level fusion of fingerprint and finger-vein for personal identification," *Pattern Recognit. Lett.*, vol. 33, no. 5, pp. 623–628, Apr. 2012.
[23] K.-H. Pong and K.-M. Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognit.*, vol. 47, no. 2, pp. 556–567, Feb. 2014.
[24] W.-P. Li, J. Yang, and J.-P. Zhang, "Uncertain canonical correlation analysis for multi-view feature extraction from uncertain data streams," *Neurocomputing*, vol. 149, pp. 1337–1347, Feb. 2015.
[25] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Fully automatic face normalization and single sample face recognition in unconstrained environments," *Expert Syst. Appl.*, vol. 47, pp. 23–34, Apr. 2016.
[26] S. Setumin and S. A. Suandi, "Difference of Gaussian oriented gradient histogram for face sketch to photo matching," *IEEE Access*, vol. 6, pp. 39344–39352, 2018.
[27] J. Sommer. (2015). *10 Incredibly Realistic Sketches By the World's Most Successful Forensic Artist*. Accessed: Oct. 10, 2016. [Online]. Available: http://www.businessinsider.com/10-sketches-by-forensic-artist-lois-gibson-2015-7
[28] R. G. Uhl and N. da Vitoria Lobo, "A framework for recognizing a facial image from a police sketch," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1996, pp. 586–593.
[29] 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.
[30] M. Zhu, N. Wang, X. Gao, and J. Li, "Deep graphical feature learning for face sketch synthesis," in *Proc. 26th Int. Joint Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2017, pp. 3574–3580.
[31] N. Wang, X. Gao, L. Sun, and J. Li, "Bayesian face sketch synthesis," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1264–1274, Mar. 2017.
[32] B. Cao, N. Wang, X. Gao, and J. Li, "Asymmetric joint learning for heterogeneous face recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2018, pp. 6682–6688.
[33] A. Radman and S. A. Suandi, "Robust face pseudo-sketch synthesis and recognition using morphological-arithmetic operations and HOG-PCA," *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 25311–25332, 2018.
[34] 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.
[35] L. Jiao, S. Zhang, L. Li, F. Liu, and W. Ma, "A modified convolutional neural network for face sketch synthesis," *Pattern Recognit.*, vol. 76, pp. 125–136, Apr. 2018.
[36] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 627–634.
[37] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 191–204, Jan. 2013.
[38] C. Galea and R. A. Farrugia, "Forensic face photo-sketch recognition using a deep learning-based architecture," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1586–1590, Nov. 2017.
[39] D. Liu, N. Wang, C. Peng, J. Li, and X. Gao, "Deep attribute guided representation for heterogeneous face recognition," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 835–841.
[40] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, 2015, p. 6.
[41] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.
[42] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," 2018, *arXiv:1801.07698*. [Online]. Available: http://arxiv.org/abs/1801.07698
[43] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1679–1686.
[44] S. Setumin and S. A. Suandi, "Cascaded static and dynamic local feature extractions for face sketch to photo matching," *IEEE Access*, vol. 7, pp. 27135–27145, 2019.
[45] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, Feb. 2017.
[46] C. Peng, X. Gao, N. Wang, and J. Li, "Sparse graphical representation based discriminant analysis for heterogeneous face recognition," *Signal Process.*, vol. 156, pp. 46–61, Mar. 2019.
[47] 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.
[48] M. Haghighat, S. Zonouz, and M. Abdel-Mottaleb, "Identification using encrypted biometrics," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Cham, Switzerland: Springer, 2013, pp. 440–448.
[49] 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.
[50] L. Shen, L. Bai, and M. Fairhurst, "Gabor wavelets and general discriminant analysis for face identification and verification," *Image Vis. Comput.*, vol. 25, no. 5, pp. 553–563, May 2007.
[51] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognit.*, vol. 24, no. 12, pp. 1167–1186, Jan. 1991.
[52] T. P. Weldon, W. E. Higgins, and D. F. Dunn, "Efficient Gabor filter design for texture segmentation," *Pattern Recognit.*, vol. 29, no. 12, pp. 2005–2015, Dec. 1996.

[53] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
[54] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
[55] R. Maini and H. Aggarwal, "Study and comparison of various image edge detection techniques," *Int. J. Image Process.*, vol. 3, no. 1, pp. 1–11, 2009.
[56] X. Tang and X. Wang, "Face photo recognition using sketch," in *Proc. Int. Conf. Image Process.*, 2002, pp. 257–260.
[57] A. M. Martinez and R. Benavente, "The AR face database," Univ. Autònoma Barcelona Edifici O, Barcelona, Spain, CVC Tech. Rep. 24, Jun. 1998. [Online]. Available: http://www.cat.uab.cat/Public/Publications/1998/MaB1998/CVCReport24.pdf
[58] 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.
[59] 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.
[60] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized MCWLD for matching sketches with digital face images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1522–1535, Oct. 2012.
[61] 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.
[62] 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.
[63] S. Klum, H. Han, A. K. Jain, and B. Klare, "Sketch based face recognition: Forensic vs. Composite sketches," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2013, pp. 1–8.
[64] P. Mittal, M. Vatsa, and R. Singh, "Composite sketch recognition via deep network—A transfer learning approach," in *Proc. Int. Conf. Biometrics (ICB)*, May 2015, pp. 251–256.
[65] S. Ouyang, T. Hospedales, Y. Z. Song, and X. Li, "Cross-modal face matching: Beyond viewed sketches," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 210–225.
[66] 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, Oct. 2017.
[67] 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.
[68] D. Gong, Z. Li, J. Liu, and Y. Qiao, "Multi-feature canonical correlation analysis for face photo-sketch image retrieval," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 617–620.
[69] A. T. Alex, V. K. Asari, and A. Mathew, "Local difference of Gaussian binary pattern: Robust features for face sketch recognition," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 1211–1216.

SAMSUL SETUMIN (Member, IEEE) 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 the Universiti Teknologi Malaysia, in 2009. He is currently pursuing the Ph.D. degree with the Universiti Sains Malaysia. Since 2010, he has been a Lecturer with the Universiti Teknologi MARA, Malaysia. He was a Test Engineer with Agilent Technologies (M) Sdn. Bhd., and Intel Microelectronics (M) Sdn. Bhd., for a period of one year. His research interests include computer vision, image processing, and pattern recognition.

MUHAMAD FARIS CHE AMINUDIN received the B.Eng. degree (Hons.) in mechatronic engineering and the M.Sc. degree in imaging from the Universiti Sains Malaysia, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Universiti Sains Malaysia, focusing on deep learning in forensic image sketch. His research interests include deep learning, image processing, and computer vision.

SHAHREL AZMIN SUANDI (Senior Member, IEEE) received the B.Eng. degree in electronic engineering and the M.Eng. and D.Eng. degrees in information science from the Kyushu Institute of Technology, Fukuoka, Japan, in 1995, 2003, and 2006, respectively. He joined the Industries, such as Sony Video (M) Sdn. Bhd., and Technology Park Malaysia Corporation Sdn. Bhd., for a period of six years, as an Engineer. He is currently a Professor and the Deputy Dean of Research, Innovation and Industry-Community Engagement with the School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Malaysia. His current research interests include face-based biometrics, real-time object detection and tracking, and pattern classification using deep learning. He served as a Reviewer for several international conferences and journals, including *IET Biometrics*, *IET Computer Vision*, *Multimedia Tools and Applications*, *Neural Computing and Applications*, *Journal of Electronic Imaging*, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE ACCESS, and so on.