Received November 12, 2019, accepted November 27, 2019, date of publication December 2, 2019,
date of current version December 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2956980

Face Occlusion Recognition With Deep Learning in Security Framework for the IoT

LI MAO[1], **FUSHENG SHENG**[2], **AND TAO ZHANG**[1,3]

[1]Department of Computer Science and Technology, Jiangnan University, Wuxi 214122, China
[2]Naval Logistics Academy Career Education Center, Tianjin 300450, China
[3]Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, Shenzhen 518060, China

Corresponding author: Tao Zhang (taozhang@jiangnan.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702226, Grant 61602253, Grant 61876072, and Grant 61602007, in part by the 111 Project under Grant B12018, in part by the Humanities and Social Sciences Projects of the Ministry of Education under Grant 18YJC760112, in part by the Chinese Postdoctoral Science Foundation under Grant 2018M632229, in part by the Experience Design Frontier Methodology and Technology Innovation Research Project (111 Project), under Grant B18027, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20170200, in part by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources under Grant KF-2018-03-065, and in part by the Fundamental Research Funds for the Central Universities under Grant JUSRP11854 and Grant JUSRP11851.

ABSTRACT Currently, the security of the Internet of Things (IoT) has aroused great concern. Face detection under arbitrary occlusion has become a key problem affecting social security. This paper designs a novel face occlusion recognition framework in the security scene of IOT, which is used to detect some crime behaviors. Our designed framework utilizes the gradient and shape cues in a deep learning model, and it has been demonstrated to be robust for its superiority to detect faces with severe occlusion. Our contributions contain three main aspects: Firstly, we present a new algorithm based on energy function for face detection; Secondly, we use the CNN models to create deep features of occluded face; Finally, to check whether the detected face is occluded, novel sparse classification model with deep learning scheme is constructed. Statistical results demonstrate that, compared with the state of the arts, our algorithm is superior in both accuracy and robustness. Our designed head detection algorithm can achieve 98.89% accuracy rate even though there are various types of severe occlusions in faces, and our designed occlusion verification scheme can achieve 97.25% accuracy rate, at a speed of 10 frames per second.

INDEX TERMS Deep learning, security framework, IoT, face occlusion recognition.

I. INTRODUCTION

The Internet of Things (IoT) has become an important research domain, and their applications have shown their potential in recent years. In order to be successful in the commercial world, deep learning technology has been broadly used in action recognition, image recognition and computer vision fields [1], [2]. Existing IoT devices, such as ATM machines, are usually used to do some financial operations all over the countries. One important reason for this is its capacity in security application, which requires a video surveillance system to distinguish potentially dangerous users from normal users solely based on the users' face [3]–[5]. Detecting potential criminals who have their faces covered is one of the applications of our paper, and this will remind ATM users not to cover his faces with something. Up to now, there have been some developmental ATM surveillance systems. The detection of partially occluded faces has been extensively researched such as biometrics and digital crime, systems used to detect abnormal face images [6]–[15] can be extensively used in many aspects. However, although several approaches have been established to improve the surveillance function for ATM, they are still helpless when dealing with severely occluded faces.

Two main algorithms are used in the process of detecting face occlusion: head region detection and occlusion verification. While most of the face detection algorithms are focussed on partial occlusion, don't perform well for arbitrary occlusions. A few years earlier, Wen et al. tried to solve the face occlusion problem based on the Gabor filter function [6]. Considering the importance of temporal feature, spatial-temporal information has gradually become the main feature of images [7]–[9]. Some reseachers tried to use skin color information to locate face region [10], [11].

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

Interestingly, some works are concerned with the popular ''recognition by parts'' scheme, which basically aims to get accurate face location by constructing effective models for analysing other human body regions [12]–[15]. All methods above are capable of handling the partial occlusion cases by teasing out other features from the non-occluded regions of human body. Nevertheless, when severe occlusion is the case, the detection rate of these algorithms will be greatly compromised. On the other hand, in some applications of IoT, color model-based methods, contour-based methods and matching-based methods can also be viewed as different applications of face detection. Color model-based methods [16]–[18] localize face parts through extracting the color information of hair and face. These algorithms are computationally straightforward but become incompetent when the head region is massively covered. Matching-based methods [19], [20] detect head through calculating the similarity between training model and the current region. Contour-based methods [21]–[23], on the other hand, make full use of shape information to characterize the contour information of head part. Although this kind of method can process the grievous occluded cases, the computational load is accordingly heavy. In addition, working with low-resolution images is not easy. Hence, this paper aimed to determine more accurate face region by designing one more effective technical framework.

Occlusion verification approaches could be classified into three main kinds: Skin Color Area Ratio-based approaches, head elements-based approaches, and classifier-based approaches. Skin Color Area Ratio-based approaches [8]–[12], [24] analyse face occlusion activity using the ratio between detected face color pixels and head area pixels. This kind of methods has strong robustness when it comes to the problem of handling pose and alignment, but it's prone to be vulnerable to changes in illumination conditions. Facial component-based methods [9], [25], [26] relied on current status of human facial components. However, when facing low resolution images, these algorithms will fail. Classifier-based approaches [28]–[30] mainly relied on some classification models. However, few training samples became the main obstacle.

Apart from geometric facial image alignment/normalisation, robust textural facial image feature extraction methods have also been developed to enhance the performance of representation-based classification algorithms. A number of studies have demonstrated that robust image feature extraction methods can promote the performance of pattern classification significantly. Classical representation-based classification methods, such as SRC, CRC and LRC, are usually based on image intensities thus perform poorly in unconstrained scenarios. To address this issue, many robust image feature descriptors, e.g. Local Binary Patterns (LBP) [31], [32] and Gabor [33], have been used in representation-based face classification and demonstrated significant improvements in accuracy. More recently, with the great success of deep neural networks, Convolutional Neural Networks (CNN) have been proven to be very effective in extracting robust image features for a variety of image classification tasks [22], [24]–[26], [34]. However, most existing deep-learning based image classification methods just simply use the nearest neighbour classifier. This motivates us to further explore the use of deep CNN features for representation-based face classification. To this end, we extract deep CNN features from geometric aligned facial images for SRC-based face classification.

In recent years, video monitoring system provides efficient data acquisition means for scene analysis and identification, and is also an important part of smart city and urban security monitoring, the understanding and recognition of complex scene, one of the important is to realize the artificial intelligence technology, the paper is mainly to monitor abnormal bahaviors in front of ATM.

Detecting facial regions in ATM surveillance is more arduous due to the next three reasons: (1)Head detection algorithm must has strong robustness to deal with arbitrary occlusion; (2) As the human body is largely covered by ATM, the body information available for analysis is very limited, regardless of being deliberately masked or not. (3) Complex illumination conditions.

Through investigating these various methods available [1]–[5], [11], [30], [35]–[40], face image detection has been widely used in practical applications, and the acquisition of face image is very important for the verification of occlusion. In recent years, some new solutions have been proposed [41]–[50]. However, when facing severe occlusion, the performance of existing systems decreases significantly.

In accordance to the aforementioned challenges, we propose a robust face occlusion recognition algorithm, which consists of two main modules: energy-based head detection and sparse classification model-based face occlusion verification. First, a new algorithm based on energy function is designed to detect elliptical head contour. It overcomes the problems common to approaches that rely on clear facial features. Then, we develop a robust face tracking framework, which uses deep regression network structure. Lastly, to confirm an occluded head region, a sparse classification model is added to our algorithm. This helps to address various skin color problems.

Our proposed algorithm mainly consists of three steps, i.e., face detection, face tracking and face verification: (1) Gaussian energy-based head localization; (2) In view of the limited training samples, we put forward an effective dictionary learning algorithm with deep features extracted from CNN in face data.(3) Sparse classification model-based face occlusion verification. A diagram elucidating the flowchart of implementation is presented in Fig 1. Firstly, the head and shoulder is located according to applying the algorithm based on gaussian energy. Secondly, an robust and fast head tracker combining with the deep network model is used to determine the face region. Lastly, a modified sparse classification model by integrating skin color and shape features is used for face occlusion verification.

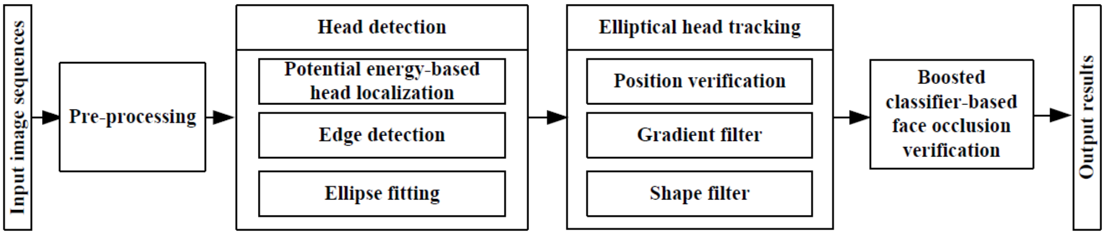

FIGURE 1. The flowchart of the proposed algorithm.

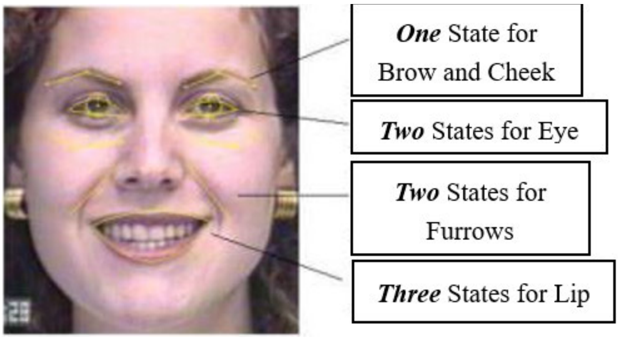

FIGURE 2. Examples of the face detection step using contour-based algorithms.

Structure of our paper is organized as follows. Section 2 introduce some works similar our proposed algorithm. Section 3 gives our designed head detection and tracking algorithm. Face occlusion verification scheme is depicted in Section 4. Section 5 gives experimental results and analysis. Finally, Section 6 gives conclusions and future works.

II. RELATED WORKS

In this section, we mainly introduce some methods similar to our proposed algorithm.

A. CONTOUR-BASED ALGORITHM

At present, the head detection with respect to contour is the only one most feasible scheme under the condition of fully occlusion, and the head detection is a complicated and challenging pattern problem, there are two main difficulties: on the one hand, it is caused by the different characteristics of head inside, head contains quite complex details, such as hair color, skin color, glasses, head ornaments and so on; on the other hand, due to the change of external conditions, the different imaging angles lead to multiple postures of the head, such as front side, side face and different features. In addition, changes in brightness, contrast and shadow of the image is also a major obstacle. All these factors make it difficult to solve the problem of head detection.

Some examples of the face detection step using contour-based algorithms are shown in Fig 2. To maintain the background of an input face image, some states are added to the obtained more landmarks for piece-wise affine warp, as shown in the figure. However, this kind of method depends on the location of the face elements and owns a high computation complexity. Long years ago, head region detection was predominantly occupied with gradient-based histograms [15]–[29], [54]–[56], but now convolutional neural networks have captured the market [51]–[53], such that more and more people are ignoring contour-based algorithms. However, there exist many aspects to be further developed or explored for these contour-based approaches.

As is presented in [54]–[57], a contour is defined as an outline that represents or surrounds the shape or form of a target in the image. Generally speaking, this kind of method is based on a priori knowledge, and there is no strict mathematical specification.Contour is closely related to edges and boundaries, which indicate the discontinuities in luminosity, geometry, and physical properties of the object in the corresponding space. It can be inferred that a clear definition about above three key factors facilitates the selection of features.

B. ORIGINAL SRC MODEL

During the past decades, sparse representation has drawn extensive attention in a variety of signal processing and image analysis applications, e.g. signal encoding, signal processing, image compression, feature representation, video analysis and image classification [16]–[35]. For image classification, the seminal work is Sparse-Representation-based Classification (SRC). In this section, we introduce the basics of the classical Sparse-Representation-based Classification (SRC) method, which is the foundation of the proposed framework, where the entire training samples consists of dictionaries that come from query facial images. The query image was categorized according to evaluation of which class gave rise to the minimal error of reconstructing it.

In original model, we define K as classes of subjects, $A = [A_1, A_2, \ldots, A_3]$ is the dictionary formed by A_i, where $A_i (i = 1, 2, \ldots, K)$ is the subset of training samples with respect to i. y represent these testing samples. Classical SRC algorithm follows the steadfast procedures.

(1) Normalize each training sample $A_i, i = 1, 2, \ldots, K$.

(2) Objection function determination and solve the l_1-minimization: $\hat{x} = \arg\min_x \left\{ \|y - Ax\|_2^2 + \gamma \|x\|_1 \right\}$, where γ is a scalar constant.

(3) Give the accurate prediction about test sample y: $Label(y) = \arg\min_i \{e_i\}$, it is a minimization process about e_i, where $e_i = \|y - A_i \hat{\alpha}_i\|_2^2$, $\hat{\alpha}_i$ standing for the coefficient vector associated with class i.

From above definition, it can be easily found that this scheme is based on the potential assumption, this assumption indicates a test sample can be reconstructed by a weighted linear combination of just those training samples attached to the same class. It is reported in [23] that it achieved impressive performance,which presented that sparse representation is naturally discriminative.

C. CLASS-SPECIFIC DICTIONARY LEARNING

Class-specific dictionary learning is defined like this: the elements in the learned dictionary $D = [D_1, D_2, \ldots, D_K]$ indicate class label correspondences to the targets classes, where D_i is the sub-dictionary with respect to class i. Our main goal is to solve representation vector $\hat{\alpha} = [\hat{\alpha}_1; \hat{\alpha}_2; \ldots; \hat{\alpha}_K]$ term, then we can use the class-specific representation residual $\left\| y - D_i\hat{\alpha}_i \right\|_2$ as classification criteria. The sub-dictionary D_i could be got through the algorithm presented in reference [33]: $\arg\min_{D_i,Z_i}\{\|A_i - D_iZ_i\|_F^2 + \lambda \|Z_i\|_1\}$, where Z_i indicates the representation matrix of A_i with respect to D_i. It's basically a basic model of learning a class-specific dictionary.

One thing to note is that, the mentioned model trains the sub-dictionaries of designed model separately, the relationship between different sub-dictionaries from different classes is ignored.

III. PROPOSED OCCLUDED FACE DETECTION SCHEME

Head detection is the most important part of our proposed system. In this part, we will show our recommended solution in ATM surveillance of IoT. We present a novel energy approximation strategy and solve the problem of face occlusion well.

A. PROPOSED GAUSSIAN ENERGY FUNCTION

Head-shoulder region shows a resemblance to an Omega (Ω), which is very much like a Gaussian curve. Illuminated by this phenomenon, we get sidetracked, maybe some appropriate Gaussian curve could approximate the head edge. By adding suitable energy terms, it is possible to reach a local minimal energy.

Denote $\Omega \in R^2$ as a closed region. The set of defined Gaussian curves is denoted as C. In addition, we introduce a scalar Lipschitz continuous function $\phi : \Omega \rightarrow R$:

$$\begin{aligned} \phi(x,y) > 0\,, &\quad (x,y) \in \Omega \\ \phi(x,y) > 0, &\quad (x,y) \notin \Omega \end{aligned} \tag{1}$$

Then, the corresponding voting points can be defined as:

$$\alpha = w(2f(x,y) - 1) \tag{2}$$

The pixels obtaining a vote are then defined by:

$$\alpha \,|w \in [-m, m] \tag{3}$$

$f(x,y)$ indicates the gray value of corresponding points.

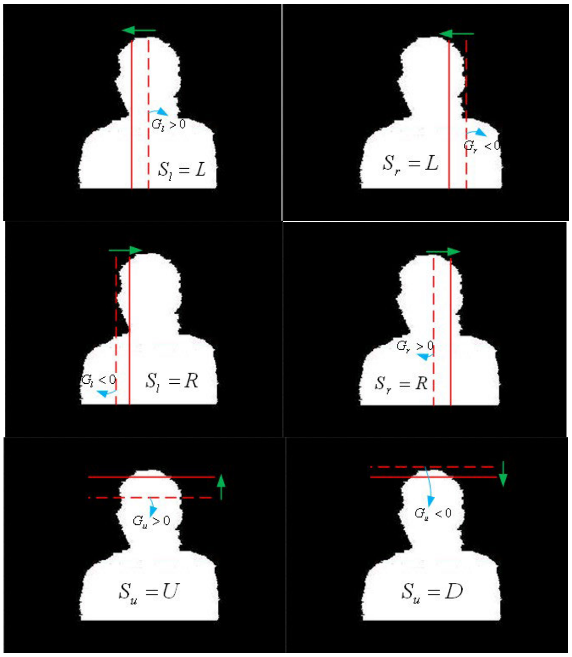

FIGURE 3. Schematic diagram of our movement rule.

Then, the potential energy function on a defined curved line C is expressed as:

$$G = \int_C \phi(x,y)g(x,y)dxdy \tag{4}$$

where $g(x,y)$ is a Gaussian distribution function.

The movement rule of the defined curved lines, i.e., the left scanning line S_l, the right scanning line S_r, and the upper scanning line S_u, is defined as:

$$\begin{aligned} &S_l = L \text{ and } S_r = L \quad \text{if } G_l > 0 \text{ or } G_r < 0 \\ &S_l = R \text{ and } S_r = R \quad \text{if } G_l < 0 \text{ or } G_r > 0 \\ &S_u = U \quad \text{if } G_u > 0 \\ &S_u = D \quad \text{if } G_u < 0 \end{aligned} \tag{5}$$

where G_l, G_r, G_u are the corresponding energy value on the left, right and upper scanning line respectively. $S_l = L$ and $S_r = L$ indicate that these lines move to the left, $S_l = R$ and $S_r = R$ indicate that these lines move to the right, $S_u = U$ describes these lines that move upward, and $S_u = D$ describes these lines that move downward. As described in Fig 3, red dotted lines indicates previous position of scanning line, while red solid lines indicate current position.

In the end, based on the above analysis, we define the following objection function:

$$T = \arg\min |G| + \beta \log(1 + |w|) \tag{6}$$

where β is a constant value that controls the trade-off between the minimal energy and fitting goodness.

Based on the aforementioned design, we construct a head detection algorithm based on Gaussian energy, which constitutes three phases: (1) Position initialization; (2) Minimizing

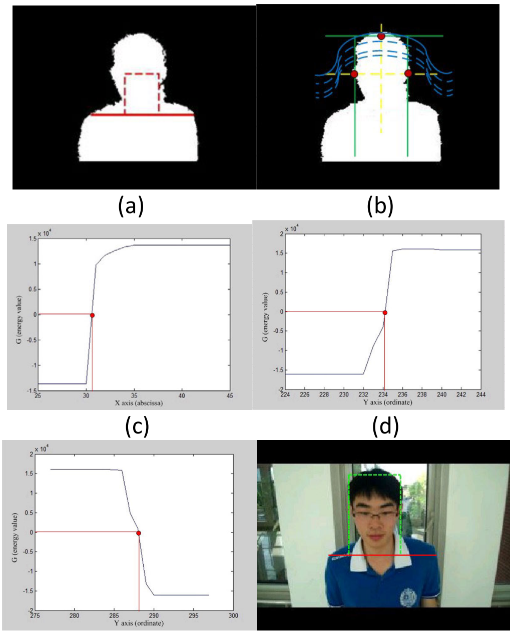

FIGURE 4. Schematic diagrams of proposed Gaussian energy-based function. (a) Initial position. (b) An example of our proposed algorithm approaching the head edge. (c) G values range when approaching the upper head edge. (d) G values range with respect to the left head edge. (e) G values range with respect to the right head edge. (f) Expected result.

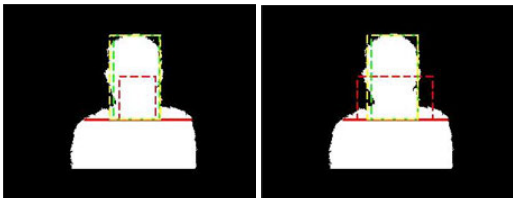

FIGURE 5. Head location results in different initial condition.

a potential energy-based iterative process; (3) Energy constraint conditions. Some more details are given in Fig 4. Also, our method is robust to any position initialization, as is shown in Fig 5.

Fig 6 shows some results of head location in a real ATM scene. As it shows, although these images are with side view, the heads' left, right and upper boundaries can still be located accurately. Our designed energy loss function is very appropriate for side view. Through such an extreme example, we can draw that our designed algorithm is very robust to change of perspective.

B. ELLIPSE FITTING

Nevertheless, our proposed method cannot get the accurate lower jaw position. Four examples are shown in Fig 7.

To handle above problem, more preprocessing will be used in this experiment. We first utilize the Canny operator to get the head edge. After that, we try to eliminate the messy edges in the inner and upper area. We then adapt the least square method to get an elliptical head region. By this means, we can precisely get the lower jaw position. Two examples illustrating this process for both a non-occluded face and an occluded face are shown in Fig 8.

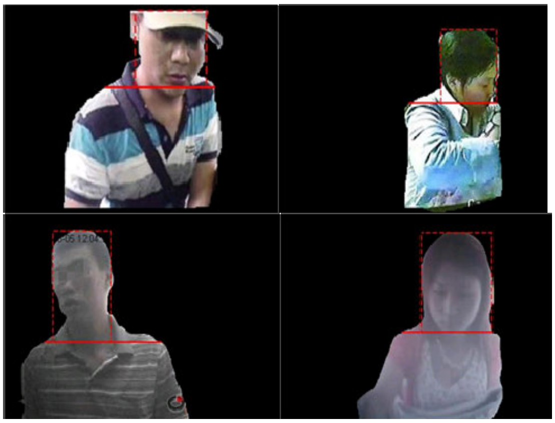

FIGURE 6. Head location results (indicated in red boxes) in sideview condition in a real ATM scene.

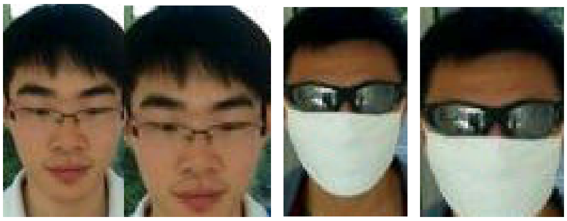

FIGURE 7. Cropped head areas located with our proposed Gaussian energy-based algorithm.

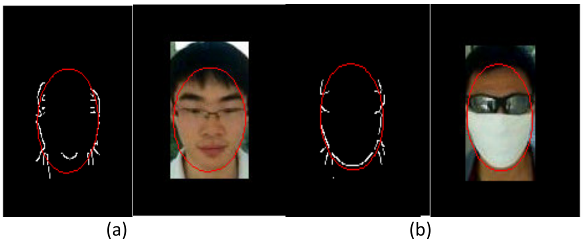

FIGURE 8. Examples of ellipse fitting results for a non-occluded face (a) and an occluded face (b). Fitted ellipse.

IV. PROPOSED OCCLUDED FACE DETECTION ALGORITHM

To verify whether the human face is covered or not, we employ a novel sparse classifier, assuming one kind of feature only, i.e., the deep CNN features.

Machine learning methods are commonly applied in image recognition problems due to their capacity to improve performance by learning from more labeled samples. With regard to the face recognition target, features that have been deeply learned need not only to be separable but also to be discriminative. Because it is difficult to collect enough labeled data, we need to build a deep model that depends on less samples. Consider enhancing the discriminative ability of the deep model, Wen etc. proposed new supervised pattern, named it as center loss. In practical application, the center is lost while learning the center information of the deep feature of each class, related report can be seen in [30]. Encouragingly, their CNNs implement the most advanced accuracy. As a result, we employ this kind of approach here to get enough discriminant features.

The classical sparse representation classification (SRC) algorithm achieves good performance in many pattern recognition tasks [31]–[37], [46]–[50]. Original and detailed explanation about SRC derives from these works [34], [35], [38]. SRC method has been applied well in some computer vision fields [35], it has aroused people's great attention. However, how to learn more effective dictionaries is still an open issue.

In order to learn more effective features, we construct an improved dictionary learning framework. While traditional SRC mainly consider the reduction of reconstruction error, our constructed model is additionally equipped with two constraints, namely the similarity constraint and coefficient incoherence constraint. The similarity constraint is utilized to seek correlations between similar descriptors through a shared dictionary space that each descriptor is projected. The coefficients incoherence constraint is used to confer poor reconstruction ability for training samples on the class-specific sub-dictionaries. Thus, both the representation residual and the constructed dictionary become more discriminative, and final classification algorithm is also built on the basis of the two items.

A. PROPOSED SPARSE CLASSIFICATION MODEL

In this paper, two parts, i.e., the similarity constraint and the coefficients incoherence constraint, are introduced to make sure the dictionaries we learned own more discriminating characteristics. The similarity constraint is utilized to seek correlations between similar descriptors through a shared dictionary space that each descriptor is projected. The coefficients incoherence constraint is used to confer poor reconstruction ability for training samples on the class-specific sub-dictionaries.

In our constructed model, dictionary D $= [D_1, D_2, \cdots, D_K]$ represent the class labels, D_i indicates the sub-dictionary of corresponding class i. In order to get more discrimative feature, we constructed a set of deep learning feature $\left\{a_{ij} | i = 1, 2, \cdots, k; j = 1, 2, \cdots, N\right\}$, where a_{ij} represents the j-th sample with respect to class i, the number of classes is denoted as K, and N indicates the total number of training samples. Let A $= [A_1, A_2, \cdots, A_i] \in R^{n\times N}$, where $A_i = [a_{i1}, a_{i2}, \cdots, a_{iN}]$, and the feature dimension is denoted as n. Our target is to incorporate the similarity constraint and coefficient incoherence constraint into the objective function of our constructed model so that the dictionary is more suitable for classification task. In addition, we build sparse code $Z = [Z_1, Z_2, \cdots, Z_i]$, D $= [d_1, d_2, \cdots, d_k] \in R^{n\times k}$ ($k > n$ and $k < N$) represents the learned dictionary. Thus, we try to construct a novel dictionary learning model:

$$\begin{aligned}\langle D, W, Z\rangle = \arg\min\{&\|A - DZ\|_F^2 + \lambda_1 \|Z\|_1 \\ &+ \lambda_2 \|Z - m\|_F^2 + \gamma_1 \|WZ - B\|_F^2 + \gamma_2 \|W\|_F^2\} \\ &s.t.\ \|d_c\|_2 \le 1, \quad \forall c \in \{1, 2, \cdots, k\}\end{aligned} \tag{7}$$

where $m = [m_1, m_2, \cdots, m_i] \in R^{k\times N}$, m_i represents mean vector Z_i, middle term $\|WZ - B\|_F^2$ represents the classification error, $B = [0, 0, \cdots, b_N] \in R^{m\times N}$ indicate the class information of input features. $b_i = [0, 0, \cdots 1 \cdots, 0]^T \in R^m$ indicates corresponding vector. $W \in R^{m\times k}$ is the parameters matrix, and λ_1, λ_2, γ_1 and γ_2 represent the normalized regulatory factors.

In the constructed learning scheme, the similarity constraint $\|WZ - B\|_F^2$ and coefficients incoherence constraint $\|W\|_F^2$ are designed in Eq.7.

Similarity constraint is utilized to seek correlations between similar descriptors through a shared dictionary space that each descriptor is projected. In our constructed dictionary learning model, W constitute with a series of classifier parameters, thus, we can obtain the minimum classification error through minimizing $\|W\|_F^2$.

Overall, minimizing the similarity constraint ensures that learned dictionary has better representation form of corresponding samples and minimizing the coefficients incoherence constraint enforces a confer poor reconstruction ability for training samples on the class-specific sub-dictionaries. By fusing the similarity constraint and coefficients incoherence constraint, the constructed dictionary become more discriminative for the classification task.

B. OPTIMIZATION OF PROPOSED SPARSE MODEL

It's easy to observe that the function of Eq. 7 is a solving co-convex problem. Generally speaking, we can solve this optimization problem by assuming that two variables are constant value.In the following part, we will introduce the specific solution process in detail.

Updating Z: At this time, D and W are considered to be two known values, our main task is to find a solution to Z. It's worth noting that, if Z_i is updated, all $Z_j (j \neq i)$ can be considered to be constant. Therefore, we get the following solution:

$$\begin{aligned}\langle Z\rangle = \arg\min\{&\|A - DZ\|_F^2 + \lambda_1 \|Z\|_1 + \lambda_2 \|Z - m\|_F^2 \\ &+ \gamma_1 \|WZ - B\|_F^2\end{aligned} \tag{8}$$

And if we solve it further, we can get:

$$\begin{aligned}Z_i = &\left\{D^T D + (\lambda_1 + \lambda_2)I + \gamma_1 W^T W\right\}^{-1} (D^T A_i + \lambda_2 m_i \\ &+ \gamma_1 W^T b_i)\end{aligned} \tag{9}$$

Updating D: At this time, Z and W are considered to be two known values, our main task is to find a solution to D. If D_i is updated, all $D_j (j \neq i)$ can be considered to be constant. Therefore, Eq. 7 is further deformation into:

$$\begin{aligned}\langle D\rangle = \arg\min &\left\{\|A - DZ\|_F^2\right\}, \\ &s.t.\ \|d_c\|_2 = 1, \quad \forall c \in \{1, 2, \cdots, k\}.\end{aligned} \tag{10}$$

The above problem in Eq. 10 is usually solved by classical Lagrange dual scheme.

Updating W: At this time, Z and D are considered to be two known values, our main task is to find a solution to W, Eq. 7 is further deformation into:

$$\langle W \rangle = \arg\min \left\{ \gamma_1 \left\| WZ - B \right\|_F^2 + \gamma_2 \left\| W \right\|_F^2 \right\} \tag{11}$$

It's easy to observe that the function of Eq. 11 is a solving least square problem. Finally, we can obtain:

$$W_i = b_i Z_i^T (Z_i Z_i^T + \frac{\gamma_2}{\gamma_1} I)^{-1} \tag{12}$$

Therefore, on the basis of the above solution process, all parameters of Eq. 7 are easy to be obtained.

C. PROPOSED SPARSE MODEL-BASED CLASSIFIER

Based on the parameters obtained from the above solutions, we started to design the classification task based on the learned dictionary. Considering the different classification tasks, we try to use different strategies.

The following is the representation model we have designed:

$$\hat{\alpha} = \arg\min \left\{ \left\| y - D\alpha \right\|_F^2 + \gamma \left\| \alpha \right\|_2 \right\} \tag{13}$$

where γ is a constant value, and $\hat{\alpha} = [\hat{\alpha}^1, \hat{\alpha}^2, \cdots, \hat{\alpha}^k]^T$, where $\hat{\alpha}^i$ represents the sub-vector with respect to dictionary D_i. Based on the similarity constraint and coefficients incoherence constraint, the constructed dictionary become more discriminative for the classification task, which can be described as follows:

$$l = W\hat{\alpha} \tag{14}$$

It's easy to observe that Eq. 14 is a linear predictive classifier, our goal is mainly to seek the label indexes, which are closely related to the largest element of above vector l.

V. EXPERIMENTAL RESULTS

Considering the face occlusion detection framework constructed in our paper for arbitrarily occluded heads under complex background, we tested a large quantity of video sequences through using a common and stationary DVR (www.pami.sjtu.edu.cn/people/zhangtao/), which simulate the monitoring system of banks. Detailed verification was carried out by analysing 120 video sequences of 8 objects (40 video sequences for four cases: no occlusion, mouth occlusion, head occlusion and other face occlusion). Experimental results of face occlusion and head detection are compared with the latest algorithms respectively.

The model we designed above can be divided into dictionary learning phase and constructed classifier phase. In the process of solving the dictionary parameters, we have $\lambda_1 = 0 : 005$; $\lambda_2 = 3$; $\gamma_1 = 1$; $\gamma_2 = 0.1$; and in classification phase, we set $\gamma = 0.01$.

A. PRE-PROCESSING

Compared to the traditional foreground extraction algorithm [51]–[60], the frame difference method has an obvious advantage of low computational cost, which motivates us to use a frame difference method to detect moving objects. However, this kind of methods has three disadvantages: 1) The appropriate threshold of frame difference is difficult to select, i.e., due to the noise, a inappropriate value can lead to false edges; 2) It can not detect still objects, i.e., objects with little or no motion within a short period; 3) The selection of appropriate frame interval that produces enough object motion is difficult. In order to solve these problems, we improve it mainly from the following two aspects: the three-sigma rule [61], and the kurtosis-based frame selection.

TABLE 1. Performance comparison on face occlusion detection between our algorithm and other method.

Algorithms	Hit	HitO	AFA	Time
Ours	**98.89%**	**97.25%**	**1.86%**	350ms
Jia et al [19]	86.97%	95.62%	2.31%	431ms
Zhou et al [20]	86.67%	96.53%	2.09%	418ms
Dong et al [10]	64.29%	96.07%	2.77%	293ms
Sharma et al [40]	66.06%	96.18%	3.02%	**255ms**
Tan et al [14]	83.09%	95.82%	2.62%	298ms
Oh et al [17]	84.47%	96.18%	3.09%	437ms

A fast reference frame selection algorithm was proposed by Pan *et al.* [39] to make the numerous reference frames computationally straightforward for foreground extraction. For the above idea, we extract foreground by using the frame interval selection and the adaptive thresholding method based on kurtosis in [61], and afterwards both the calculation result of frame difference method and the edge detection image are considered to generate more complete motion edge. On the other hand, the approach in [28] is used to to eliminate the effects of light. In order to fill the holes and eliminate some small areas caused by uneven illumination, image erosion and image denoising operations are performed. Then, a horizontal filling algorithm is applied, which scans each line of foreground and pixels between the starting point and end point are all set as foreground points. Finally, we can get a filled foreground image.

B. EXPERIMENTAL RESULTS

In this part, we have done a lot of verification on the collected video data based on the face occlusion verification algorithm proposed by us.

In order to verify the excellent performance of our constructed algorithm quantitatively. This paper makes a detailed analysis between our algorithm with the state-of-the-art methods [10], [14], [17], [19], [20], [40] implemented by us. Table 1 gives the comparation of the detection rate of head (Hit), the detection rate of face occlusion verification (HitO), the average false alarm rate (AFA) and the average processing time. The classification performance is also evaluated in terms of average accuracy and standard deviation (acc ± std) in Table 2. From the results shown in following two tables, it is evidenced that our proposed algorithm achieved best preformance with the state of the art.

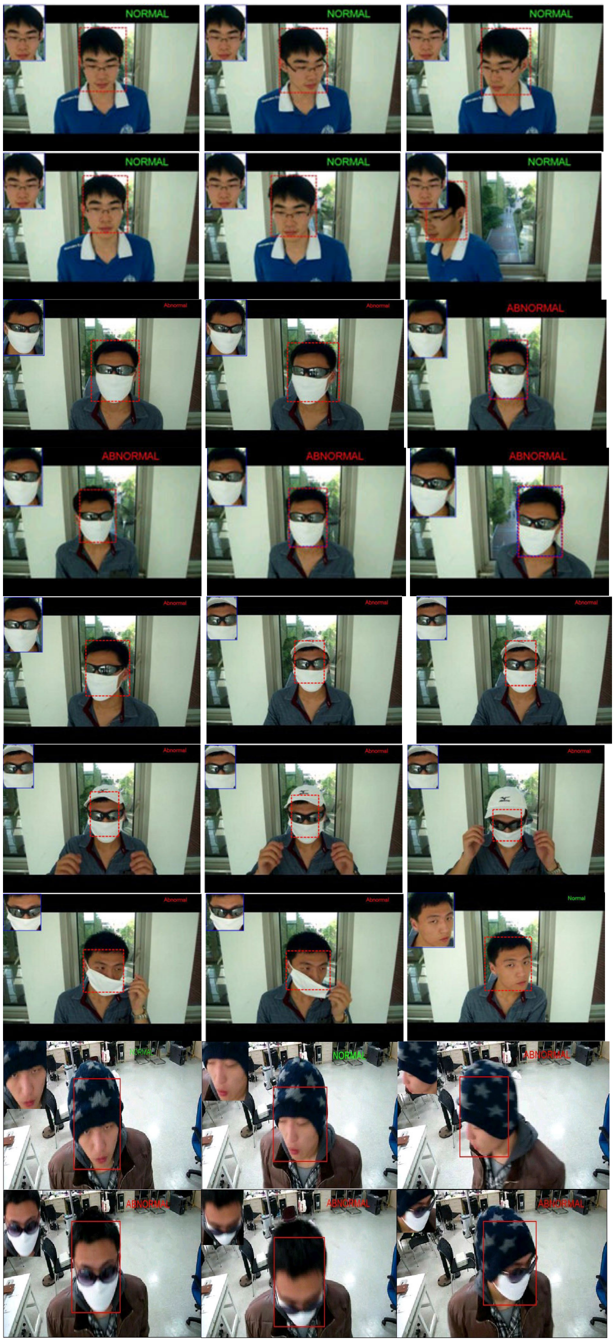

FIGURE 9. Results of occluded face detection.

Our method perfectly outperforms all others in all aspects. Moreover, by trialing a large amount of test campaigns and more realistic scenes, our evaluation result is more credible. Our constructed system works at an average speed of 2-5 frames per second. Our solution outperforms state of the art methods with the highest accuracy rates and strongest practicability in all three cases. And for practical consideration, we alarm when occluded face is detected in consecutive several frames. Fig 9 gives the examples of detection results.

To verify the statistical significance, the classification performance is tested in view of average accuracy and standard deviation (acc ± std) the state-of-the-art methods [19], [20], [40], [57]–[59], as is shown in Table 2. Comparing with the recent approaches, our proposed classification framework perform better. In most cases, the precision of our approach is at least 2% better than the others, except segment-based method [58] and low-rank regression method [59]. To give more statistical analysis, we perform *t-test* for the precision obtained by segment-based method and low-rank regression method and by our approach on our built ATM dataset, under the null hypothesis using a significance level of 0.05. *p-value* is found as 0.00082, indicating that accuracy rate performed by our approach is indeed significantly better than segment-based method and low-rank regression method.

TABLE 2. Comparisons of classification results(Accuracy:%).

Algorithms	acc ± std
Ours	**96.39 ± 1.47**
Jia et al [19]	94.19 ± 1.87
Zhou et al [20]	94.35 ± 1.76
Sharma et al [40]	94.28 ± 1.81
Ding et al [57]	94.59 ± 1.70
Mahbub et al [58]	95.14 ± 1.58
Wu et al [59]	94.91 ± 1.64

VI. CONCLUSION

In this paper, we focus on developing a real-time occluded face detection method. Different from the most existing face detection algorithms aiming at detecting a normal and unshaded faces, we have constructed an effective occluded face detection mechanism which has better accuracy in the detection based on the reduced computational complexity. The background of this paper is surveillance of IoT, which is mainly applied in target detection and recognizing at ATM intelligent surveillance system. In this paper, we proposed a novel face occlusion recognition framework. It provides three stages, i.e., face detection, face tracking and face verification. A novel energy with Gaussian curve is proposed, and then novel face tracking algorithm with deep CNN feature we applied, finally, we design a novel dictionary learning scheme for verification task.

Our proposed head detection algorithm is shown to be robust to detect faces with arbitrary poses. In addition, the face occlusion verification algorithm we proposed can effectively distinguish whether the face area is occluded or not. Experimental results indicate that our designed head detection algorithm can achieve 98.89% accuracy rate even though there are various types of severe occlusions in faces, and our designed occlusion verification scheme can achieve 97.25% accuracy rate, at a speed of 10 frames per second.

The vision-based face occlusion detection is a relatively independent part that touches wider areas, and there exist many problems that have not been solved perfectly. In future work, we are planning to focus on the following issues:

1) To develop other head detectors which are more capable of dealing with a variety of head poses. For example, in some cases where the angle of tilt is very large. 2) Find these algorithms that are capable of detecting faces for all races, sexes, and ages. For example, Westerners, Europeans, Africans. 3) Do some improvements so that the algorithm will not fail to detect these faces in the cases of severe variations in skin color due to large illumination changes. Extreme lighting conditions can be fatal. 4) To facilitate the research of other scholars, building more ground-truth datasets with respect to ATM surveillance scene is necessary.

REFERENCES

[1] P. M. Kumar, U. Gandhi, R. Varatharajan, G. Manogaran, J. R, and T. Vadivel, "Intelligent face recognition and navigation system using neural learning for smart security in Internet of Things," *Cluster Comput.*, vol. 22, pp. 7733–7744, 2017.

[2] A. Souri, A. Hussien, and M. Hoseyninezhad, "A systematic review of IoT communication strategies for an efficient smart environment," *Trans. Emerg. Telecommun. Technol.*, to be published.

[3] X. Wang, X. Wang, and S. Mao, "RF sensing in the Internet of Things: A general deep learning framework," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 62–67, Sep. 2018.

[4] A. Ferdowsi and W. Saad, "Deep learning-based dynamic watermarking for secure signal authentication in the Internet of Things," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[5] A. Souri and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques," *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, p. 3, Dec. 2018.

[6] J. Kim, Y. Sung, S. M. Yoon, and B. G. Park, "A new video surveillance system employing occluded face detection," in *Innovations in Applied Artificial Intelligence* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2005, pp. 65–68.

[7] C.-Y. Wen, S.-H. Chiu, Y.-R. Tseng, and C.-P. Lu, "The mask detection technology for occluded face analysis in the surveillance system," *J. Forensic Sci.*, vol. 50, no. 3, pp. 593–601, 2005.

[8] D.-T. Lin and M.-J. Liu, "Face occlusion detection for automated teller machine surveillance," in *Advances in Image and Video Technology* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2006, pp. 641–651.

[9] G. Kim, J. K. Suhr, H. G. Jung, and J. Kim, "Face occlusion detection by using b-spline active contour and skin color information," in *Proc. Int. Conf. Control, Autom., Robot. Vis.*, 2010, pp. 627–632.

[10] C.-M. Chen, B. Xiang, Y. Liu, and K.-H. Wang, "A secure authentication protocol for Internet of vehicles," *IEEE Access*, vol. 7, no. 1, pp. 12047–12057, 2019.

[11] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognit.*, vol. 3, no. 3, pp. 1106–1122, 2007.

[12] Y. Zhang and A. M. Martinez, "A weighted probabilistic approach to face recognition from multiple images and video sequences," *Image Vis. Comput*, vol. 6, no. 6, pp. 626–638, 2006.

[13] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ICA for face recognition robust to local distortion and partial occlusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 12, pp. 1977–1981, Dec. 2005.

[14] X. Tan, S. Chen, H. Zhou, and F. Zhang, "Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble," *IEEE Trans. Neural Netw.*, vol. 4, no. 4, pp. 875–886, Jul. 2005.

[15] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 3, no. 3, pp. 337–350, Mar. 2006.

[16] Q. Liu, W. Yan, H. Lu, and S. Ma, "Occlusion robust face recognition with dynamic similarity features," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, Aug. 2006, pp. 544–547.

[17] H. J. Oh, K. M. Lee, and S. U. Lee, "Occlusion invariant face recognition using selective local non-negative matrix factorization basis images," *Image Vis. Comput.*, vol. 11, no. 11, pp. 1515–1523, 2008.

[18] C.-M. Chen, K.-H. Wang, K.-H. Yeh, B. Xiang, T.-Y. Wu, "Attacks and solutions on a three-party password-based authenticated key exchange protocol for wireless communications," *J. Ambient Intell. Hum. Comput.*, vol. 10, no. 8, pp. 3133–3142, Aug. 2019.

[19] H. Jia and A. Martinez, "Support vector machines in face recognition with occlusions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 136–141.

[20] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using Markov random fields," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, Aug./Oct. 2009, pp. 1050–1057.

[21] J. Lin, J. Ming, and D. Crookes, "Robust face recognition with partial occlusion, illumination variation and limited training data by optimal feature selection," *IET Comput. Vis.*, vol. 5, no. 1, pp. 23–32, Jan. 2011.

[22] G. Guo, H. Wang, Y. Yan, J. Zheng, and B. Li, "A fast face detection method via convolutional neural network," *Neurocomputing*, to be published.

[23] S.-M. Huang and J.-F. Yang, "Robust face recognition under different facial expressions, illumination variations and partial occlusions," in *Proc. 17th Int. Conf. Adv. Multimedia Modeling (MMM)*, vol. 2, 2011, pp. 326–336.

[24] H. Wu, K. Zhang, and G. Tian, "Simultaneous face detection and pose estimation using convolutional neural network cascade," *IEEE Access*, vol. 6, pp. 49563–49575, 2019.

[25] Y. Liu and M. D. Levine, "Multi-path region-based convolutional neural network for accurate detection of unconstrained 'hard faces,'" in *Proc. Conf. Comput. Robot Vis.*, 2018, pp. 183–190.

[26] S. Zhang, L. Wen, H. Shi, Z. Lei, S. Lyu, and S. Z. Li, "Single-shot scale-aware network for real-time face detection," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 537–559, 2019.

[27] Z. Yao and H. Li, "Tracking a detected face with dynamic programming," *Image Vis. Comput.*, vol. 24, no. 6, pp. 573–580, 2006.

[28] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, "On the removal of shadows from images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 59–68, Jan. 2006.

[29] W. Zou, Y. Li, K. Yuan, and D. Xu, "Real-time elliptical head contour detection under arbitrary pose and wide distance range," *J. Vis. Commun. Image R.*, vol. 20, no. 3, pp. 217–228, 2009.

[30] Y. D. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 499–515.

[31] Z. Fan, M. Ni, Q. Zhu, and C. Sun, "L_0-norm sparse representation based on modified genetic algorithm for face recognition," *J. Vis. Commun. Image Represent*, vol. 28, pp. 15–20, Apr. 2015.

[32] X. Song, Y. Chen, Z.-H. Feng, G. Hu, T. Zhang, and X.-J. Wu, "Collaborative representation based face classification exploiting block weighted LBP and analysis dictionary learning," *Pattern Recognit.*, vol. 88, pp. 127–138, Apr. 2019.

[33] T. Zhang, W. Jia, X. He, and J. Yang, "Discriminative dictionary learning with motion Weber local descriptor for violence detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 696–709, Mar. 2017.

[34] T. Zhang, J. Li, W. Jia, J. Sun, and H. Yang, "Fast and robust occluded face detection in ATM surveillance," *Pattern Recognit. Lett.*, vol. 107, pp. 33–40, May 2018.

[35] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[36] Y. Xu, B. Zhang, and Z. Zhong, "Multiple representations and sparse representation for image classification," *Pattern Recognit.*, vol. 68, pp. 9–14, Dec. 2015.

[37] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[38] Z. Zhang, Y. Xu, X. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, pp. 490–530, 2015.

[39] Z. Pan, P. Jin, J. Lei, Y. Zhang, X. Sun, and S. Kwong, "Fast reference frame selection based on content similarity for low complexity HEVC encoder," *J. Vis. Commun. Image Represent.*, vol. 40, pp. 516–524, Oct. 2016.

[40] M. Sharma, S. Prakash, and P. Gupta, "An efficient partial occluded face recognition system," *Neurocomputing*, vol. 116, pp. 231–241, Sep. 2013.

[41] G. Lin, Y. Liu, and W. Zhu, "Speeding up a memetic algorithm for the max-bisection problem," *Numer. Algebra Control Optim.*, vol. 5, no. 2, pp. 151–168, 2017.

[42] Z. R. Shi, P. P. Lee, J. Shu, and W. Guo, "Encoding-aware data placement for efficient degraded reads in XOR-coded storage systems: Algorithms and evaluation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 12, pp. 2757–2770, Dec. 2018.

[43] J. Chen, Y. Liu, Z. Zhu, and W. Zhu, "An adaptive hybrid memetic algorithm for thermal-aware non-slicing VLSI floorplanning," *Integr., VLSI J.*, vol. 58, pp. 245–252, Jun. 2017.

[44] Y. Niu, J. Chen, and W. Guo, "Meta-metric for saliency detection evaluation metrics based on application preference," *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 26351–26369, 2018.

[45] X. Li, Z. Zhu, and W. Zhu, "Discrete relaxation method for triple patterning lithography layout decomposition," *IEEE Trans. Comput.*, vol. 66, no. 2, pp. 285–298, Feb. 2017.

[46] L. Guo and H. Shen, "Efficient approximation algorithms for the bounded flexible scheduling problem in clouds," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 12, pp. 3511–3520, Dec. 2017.

[47] Y. Cheng, F. Wang, H. Jiang, Y. Hua, D. Feng, L. Zhang, and J. Zhou, "A communication-reduced and computation-balanced framework for fast graph computation," *Frontiers Comput. Sci.*, vol. 12, no. 5, pp. 887–907, 2018.

[48] D. Yang, X. Liao, H. Shen, X. Cheng, and G. Chen, "Relative influence maximization in competitive social networks," *Sci. China Inf. Sci.*, vol. 60, no. 10, p. 108101, 2017.

[49] J. Wei, X. Liao, and H. Zheng, "Learning from context: A mutual reinforcement model for Chinese microblog opinion retrieval," *Frontiers Comput. Sci.*, vol. 12, no. 4, pp. 714–724, 2017.

[50] Y. Yang, X. Zheng, and C. Tang, "Lightweight distributed secure data management system for health Internet of Things," *J. Netw. Comput. Appl.*, vol. 89, pp. 26–37, Jul. 2017.

[51] M. I. Razzak and S. Naz, "Microscopic blood smear segmentation and classification using deep contour aware CNN and extreme machine learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 801–807.

[52] Z. Zhang, F. Xing, H. Su, X. Shi, and L. Yang, "Recent advances in the applications of convolutional neural networks to medical image contour detection," 2017, *arXiv:1708.07281*. [Online]. Available: https://arxiv.org/abs/1708.07281

[53] A. Y. Yang and L. Cheng, "Long-bone fracture detection using artificial neural networks based on contour features of X-ray images," 2019, *arXiv:1902.07458*. [Online]. Available: https://arxiv.org/abs/1902.07897

[54] Y. Ming, H. Li, and X. He, "Contour completion without region segmentation," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3597–3611, Aug. 2016.

[55] C. Rasche, "Rapid contour detection for image classification," *IET Image Process.*, vol. 12, no. 4, pp. 532–538, 2018.

[56] X. M. Kang, Q. Kong, Y. Zeng, and B. Xu, "A fast contour detection model inspired by biological mechanisms in primary vision system," *Frontiers Comput. Neurosci.*, vol. 12, pp. 28–39, Apr. 2018.

[57] C. Ding and D. Tao, "Pose-invariant face recognition with homography-based normalization," *Pattern Recog.*, vol. 66, pp. 144–152, Jun. 2017.

[58] U. Mahbub, S. Sarkar, and R. Chellappa, "Segment-based methods for facial attribute detection from partial faces," *IEEE Trans. Affect. Comput.*, to be published.

[59] C. Y. Wu and J. J. Ding, "Occluded face recognition using low-rank regression with generalized gradient direction," *Pattern Recognit.*, vol. 80, pp. 256–268, Aug. 2018.

[60] J.-S. Pan, C.-Y. Lee, A. Sghaier, M. Zeghid, and J. Xie, "Novel systolization of subquadratic space complexity multipliers based on toeplitz matrix-vector product approach," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 7, pp. 1614–1622, Mar. 2019.

[61] M. R. Dileep and A. Danti, "Human age and gender prediction based on neural networks and three sigma control limits," *Appl. Artif. Intell.*, vol. 32, no. 3, pp. 1–12, 2018.

LI MAO received the B.Sc. degree from Southeast University, Nanjing, China, in 1990, and the M.S. degree from Donghua University, Shanghai, China, in 2003. He is currently an Associate Professor with the Department of Computer Science, Jiangnan University. His major research interests include visual surveillance, object detection, and data mining.

FUSHENG SHENG received the bachelor's degree from Northeast Normal University, in 1995, and the master's degree from the Huazhong University of Science and Technology, in 2008. He is currently the Director of the Military Professional Educating Technology Center, Naval Logistics Academy. His major research interests include research on informatization, image processing and pattern recognition, logistical simulating training, designing and developing of software, and the research and application of the military professional education.

TAO ZHANG received the bachelor's degree from Henan Polytechnic University, China, in 2008, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China. He is currently an Associate Professor with the Department of Computer Science, Jiangnan University. His major research interests include visual surveillance, object detection, and pattern analysis.