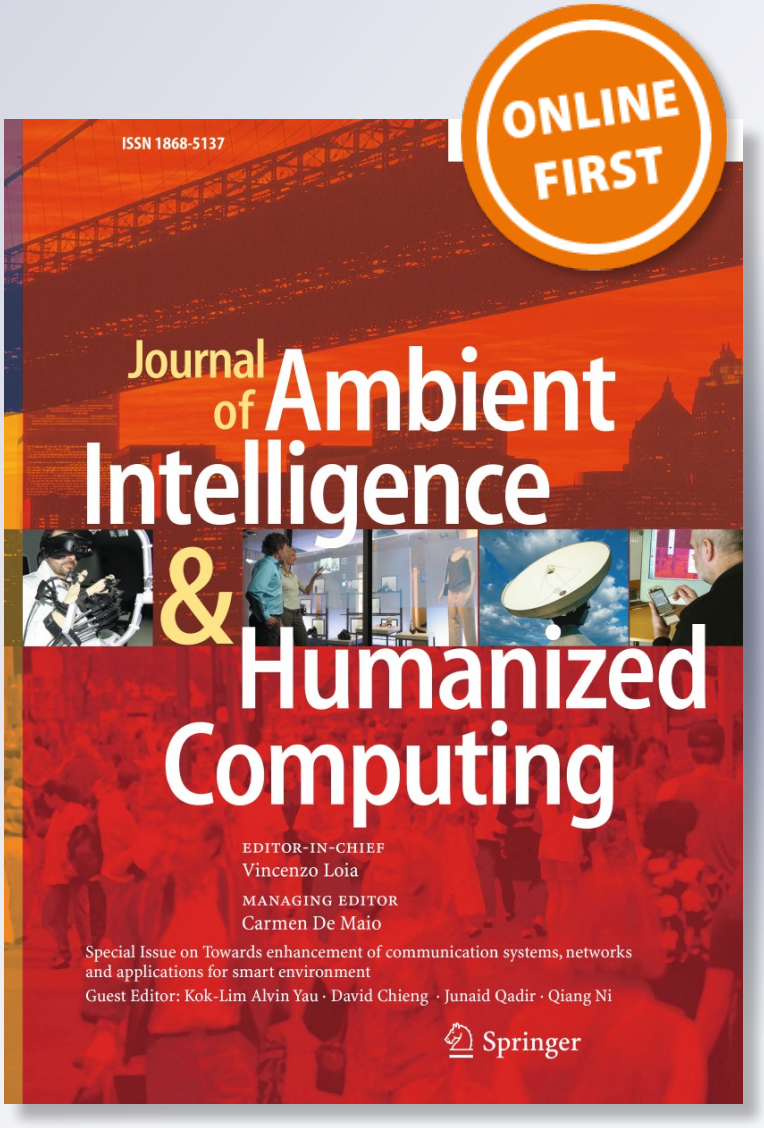
Image gradient orientations embedded structural error coding for face recognition with occlusion

Xiao-Xin Li, Pengyi Hao, Lin He & Yuanjing Feng

Journal of Ambient Intelligence and Humanized Computing

ISSN 1868-5137

J Ambient Intell Human Comput
DOI 10.1007/s12652-019-01257-7



Springer

Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-019-01257-7

ORIGINAL RESEARCH

Image gradient orientations embedded structural error coding for face recognition with occlusion

Xiao-Xin Li[1] · Pengyi Hao[1] · Lin He[2] · Yuanjing Feng[1]

Received: 15 January 2018 / Accepted: 19 February 2019

Abstract
Partially occluded faces are very common in automatic face recognition (FR) in the real world. We explore the problem of FR with occlusion by embedding Image Gradient Orientations (IGO) into robust error coding. The existing works usually put stress on the error distribution in the non-occluded region but neglect the one in the occluded region due to its unpredictability incurred by irregular occlusion. However, in the IGO domain, the error distribution in the occluded region can be built simply and elegantly by a uniform distribution in the interval $[-\pi, \pi)$, and the one in the occluded region can be well built by a weight-conditional Gaussian distribution. By incorporating the two error distributions and a Markov random field for the priori distribution of the occlusion support, we propose a joint probabilistic generative model for a novel IGO-embedded Structural Error Coding (IGO-SEC) model. Two methods, a new reconstruction method and a new robust structural error metric, are further presented to boost the performance of IGO-SEC. Extensive experiments on 8 popular robust FR methods and 4 benchmark face databases demonstrate the effectiveness and robustness of IGO-SEC in dealing with facial occlusion and occlusion-like variations.

Keywords Unconstrained face recognition · Face occlusion · Image gradient orientations · Structural error coding · Markov random field

1 Introduction

In unconstrained settings (Hua et al. 2011), facial occlusions are very common. When there exists partial occlusions or disguises in the test face images, robust face recognition systems (He et al. 2011; Yang et al. 2013c; Sun et al. 2015; Parkhi et al. 2015; Wu et al. 2015; Ghazi and Ekenel 2016; Zhao et al. 2018) might fail to perform recognition accurately. Existing works (Wright et al. 2009; Wright and Ma 2010; He et al. 2011; Li et al. 2013; Liang and Li 2015) illustrate that it is not the missing discriminative information caused by occlusion but the common high order statistical structures (localization, orientation, and bandpass) shared by occlusions and face images that mainly account for the performance drop of existing face recognition systems. Therefore, the primary task for face recognition with occlusion is to effectively suppress or eliminate the influence of occlusions.

The extant works can be mainly divided into two types: robust error coding and robust feature extraction. Robust error coding methods mainly concern how to correct or reduce the errors incurred by occlusion. One main solving clue is to build an occlusion probability estimation map according to the *error image*, also dubbed *reconstruction error*, between the occluded face and its reconstruction. Within the framework of sparse coding, the probabilistic generative model of the reconstruction error has been extensively explored (Chen et al. 2001; Wright et al. 2009; Wright and Ma 2010). These models usually assume the reconstruction error follows a specific distribution. However, recent researches (Yang et al. 2011, 2013c) showed that the reconstruction error may be far from any specific distribution, and pay attention to the weighted or separated error distribution. The models based weighted error distribution, such as Robust Sparse Coding (RSC) (Yang et al. 2011),

✉ Xiao-Xin Li
mordekai@zjut.edu.cn

[1] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, People's Republic of China

[2] School of Automation Science and Engineering, South China University of Technology, Guangzhou, People's Republic of China

Published online: 14 March 2019

CorrEntropy-based Sparse Representation (CESR) (He et al. 2011), Sparse Error Correction with Markov Random Fields (SEC-MRF) (Zhou et al. 2009), and Structured Sparse Error Coding (SSEC) (Li et al. 2013), try to weaken the effect of the error caused by outliers and focus on the error distribution in the non-occluded region. In contrast, the error distribution in the occluded region is also very important, since it indicates how the occlusions are *understood* by the model. (Tzimiropoulos et al. 2012) observed that the gradient orientation differences of dissimilar images followed a *uniform distribution* on the interval $[-\pi, \pi)$ with a high significant level. With this observation, they further showed that local orientation mismatches caused by outliers can be canceled out when the cosine kernel is applied, and proposed an **IGO** (*Image Gradient Orientations*)-embedded subspace learning framework, such as IGO-PCA and IGO-LDA.

However, no matter the error coding models concerning the error distribution in the occluded or non-occluded region, they all need to add an new hidden variable, i.e., the error weight, which to some extent is costly to be modeled and estimated. More recently, Yang et al. (Luo et al. 2015; Qian et al. 2015; Yang et al. 2017) proposed a series of nuclear norm based error coding models without using an error weight or distinguishing the occluded region from non-occluded region, as they discovered that the main features of the reconstruction error can be reflected in its singular values, which are stably sparse. Their works provide a new viewpoint to understand the reconstruction error.

Compared to the robust error coding models, robust feature extraction methods have long been underestimated for face recognition with occlusion. In (Wright et al. 2009) indicated that "feature extraction schemes would discard useful information that could help compensate for the occlusion, and no representation is more redundant, robust, or informative than the original images". Such an assertion is being challenged by recent works. Both the shallow features, such as Gabor features (Yang et al. 2013a), IGO (Tzimiropoulos et al. 2012), and AdapWeber (Li et al. 2017), and the deep features, such as PCANet (Chan et al. 2015), VGG-Face (Parkhi et al. 2015), DeepID (Sun et al. 2015), DeepFace (Taigman et al. 2015), and LCNN (Wu et al. 2015), showed comparable benefits in dealing with facial occlusion. (Sun et al. 2015) even claimed that "deeply learned face representations are sparse, selective and robust". Despite of many improvements, (Ghazi and Ekenel 2016) and (Li et al. 2017) recently showed that deep features, to some extent, are vulnerable to occlusion.

To achieve more advanced performances, a natural idea is to embed robust features into robust error coding. To this end, (Yang et al. 2013a) first incorporated Gabor features into the sparse representation classifier (SRC) (Wright et al. 2009). The IGO methods (Tzimiropoulos et al. 2012) actually combines IGO features into the subspace learning framework. The above two methods, however, do not fully use the robust features, as the existing error coding models are mainly designed in pixel domain. (Yang et al. 2013b) demonstrated how to effectively use statistical local features by using the kernel tricks. Very recently, (Zhao et al. 2018) also borrowed the feature embedding idea into the deep learning schemes and proposed a robust LSTM-Autoencoders (RLA) model, which consists of an LSTM encoder (like feature extraction) and an LSTM decoder (like error coding). Compared to the shallow methods (Yang et al. 2013a; Tzimiropoulos et al. 2012; Yang et al. 2013b), the main limitation of RLA is that it needs to see all types of facial occlusions in the training stage and its performance against never-seen occlusion remains to be tested.

In this work, we continue to study how to embed the robust features into robust error coding schemes in the uncontrolled settings, where we mainly aim to solve the following three challenging problems: (1) the number of training images is relatively small; (2) occlusions along with other variations might coexist; (3) occlusions contained in the test images are not contained in the training set. To this end, we incorporate the outlier elimination function implied in the IGO features (Tzimiropoulos et al. 2012) into a newly proposed structural error coding model, which is actually induced by the framework of the joint probabilistic generative model. The study of the probabilistic generative model in the IGO domain makes our contribution mainly reflected in the following aspects:

- An independent and non-identical distribution (i.n.d.) for the reconstruction error in IGO. In existing works (Lin and Tang 2007; Zhou et al. 2009; Yang et al. 2011; Li et al. 2013), the probabilistic error model is mainly considered in the original pixel domain and assumed to be independently and identically distributed (i.i.d.) for calculating convenience. However, the i.i.d. assumption does not always work well especially when there exists other variations except for occlusions. In the IGO domain, things become easy. We use the uniform distribution, suggested by (Tzimiropoulos et al. 2012), of the reconstruction error in the occluded region, and propose a weight-conditional Gaussian distribution of the reconstruction error in the non-occluded region. We show that these two distributions can be combined to effectively describe the holistic reconstruction error.
- An IGO-embedded Structural Error Coding (IGO-SEC) model for face recognition with occlusion. By maximizing the joint probabilistic generative model in IGO, we propose an IGO-SEC model. The structures of IGO-SEC are mainly reflected in two weights imposed on the reconstruction error. One is the occlusion support used to separate occlusion from non-occlusion, and the other is the noise weight used to weaken the influences caused by

other possible variations (such as Gaussian white noise and misalignment). Experiments show that both of the two weights are very important in dealing with face recognition with occlusion.

– Two ways to boost the error image in IGO-SEC. The error image plays a central role in IGO-SEC. To build a high-quality error image, we first propose a novel reconstruction method, which constructs a high-quality reconstruction image with few noises by first calculating the reconstruction image in the pixel domain and then transforming it into the IGO domain. Then, we introduce a novel structural error metric, which reformulate the error image with good *source-and-error-separation* ability by integrating the spatial structure and statistical information of the error image in IGO. Experiments show that the boosted error image is very important to enhance the performance of IGO-SEC.

The rest of this paper is organized as follows. Sect. 2 introduces the probabilistic generative model for occluded images, which is shown to be a convenient tool to study the problem of face recognition with occlusion. Section 3 probes into the priori probabilistic mass function (PMF) of the occlusion support by exploring its contiguous spatial structure using Markov Random Field. Section 4 deeply analyzes and formulates the conditional probability density function (PDF) of the reconstruction error in the occluded and non-occluded region in the IGO domain. By integrating the priori PMF of the occlusion support and the conditional PDF of the reconstruction error, Sect. 5 proposes the IGO-SEC model and develop an alternately iterative algorithm for its optimization. Section 5 proposes two ways to improve the error image in IGO-SEC. In Sect. 8, we verify the proposed model with extensive experiments on 4 popular face databases, and compare it with eight popular face recognition methods. Finally, we give the conclusion and discuss some future work in Sect. 9.

2 Probabilistic error framework for occluded face images

Suppose we have a set of labeled training images $A = \left[A_1, A_2, \ldots, A_K\right] \in \mathbb{R}^{m\times n}$ of K subjects, where $A_k = \left[a_1^k, a_2^k, \ldots, a_{n_k}^k\right] \in \mathbb{R}^{m\times n_k}$ is a data matrix consisting of n_k training samples from subject k and $n = \sum_{k=1}^{K} n_k$. What is critical to recognize a new occluded face image $y \in \mathbb{R}^m$ from these training images is to detect and exclude its occluded region. We denote its occlusion support by $s \in \{-1, 1\}^m$, where $s_i = -1$ indicates pixel y_i is non-occluded and $s_i = 1$ indicates pixel y_i is occluded. For convenience, we further let $\dot{\mathcal{P}} = \left\{i | s_i = -1\right\}$ and $\ddot{\mathcal{P}} = \left\{i | s_i = 1\right\}$ denote the index set of the non-occluded pixels and that of the occluded ones, respectively, and then $\mathcal{P} = \dot{\mathcal{P}} \cup \ddot{\mathcal{P}}$ corresponds to the index set of the whole image pixels.

From the viewpoint of statistics, an effective way to detect the occlusion support s is to model an effective probabilistic generative model about s and its related factors, such as its host image y and y's reconstruction image $\hat{y} \in \mathbb{R}^m$ with respect to (w.r.t.) the training set A. Specially, the generative model can be built as a joint probability density function (PDF) $p(y, \hat{y}, s)$. In (Lin and Tang 2007), Lin *et al.* also incorporated the occluding object o into the generative model and formulated the joint PDF as $p(y, \hat{y}, o, s)$. However, their model seems too complex with too many parameters. For clarity and simplicity, the error coding schemes (Zhou et al. 2009; Li et al. 2013) simplify the joint PDF to $p(e, s)$ by introducing an hidden variable $e \in \mathbb{R}^m$, where $e = y - \hat{y}$ is the reconstruction error between y and $\hat{y}$.

By rewriting $p(e, s) = p(e|s)p(s)$, we can find that the joint PDF $p(e, s)$ actually embodies the idea of "divide and conquer" and "iterative refinement". Note that the occlusion support s divides not only its host image but also the reconstruction error e into the occluded and non-occluded parts. As discussed in the introduction, although the global error e might not follow any specific distribution, the error $\dot{e} \triangleq e_{\dot{\mathcal{P}}}$ in the non-occluded region and the error $\ddot{e} \triangleq e_{\ddot{\mathcal{P}}}$ in the occluded region might follow some specific distributions, respectively. (Lin and Tang 2007) assumed that both $\dot{e}$ and $\ddot{e}$ follow the Gaussian distribution, while (Li et al. 2013) suggested that both $\dot{e}$ and $\ddot{e}$, calculated by a well designed error metric, follow the exponential distribution. Both their works embody the idea of "divide and conquer" and can be summarized as the following conditional PDF

$$\begin{aligned} p(e|s) &= p(\dot{e}, \ddot{e}|s) \\ &= p(\dot{e}|s)p(\ddot{e}|s) \\ &= \prod_{i\in\dot{\mathcal{P}}} p\left(e_i|s_i = -1\right) \prod_{i\in\ddot{\mathcal{P}}} p\left(e_i|s_i = 1\right) \\ &= \prod_{i\in\mathcal{P}} \dot{p}\left(e_i\right)^{\dot{s}_i} \ddot{p}\left(e_i\right)^{\ddot{s}_i}, \end{aligned} \tag{1}$$

where we let $\dot{s} \triangleq \frac{1-s}{2}$ and $\ddot{s} \triangleq \frac{1+s}{2}$, and define $\dot{p}\left(e_i\right) \triangleq p\left(e_i|s_i = -1\right)$ and $\ddot{p}\left(e_i\right) \triangleq p\left(e_i|s_i = 1\right)$ as the conditional PDF in the non-occluded region and occluded region, respectively.

Additionally, according to the joint PDF $p(e, s)$ and using the maximum *a posteriori* (MAP) estimation, a series of well-estimated occlusion supports and reconstruction errors could be obtained in an alternating maximization way

$$e^{(t)} = \arg\max_{e} p\left(e, s^{(t-1)}\right) \tag{2}$$

$$s^{(t)} = \arg\max_{s} p\left(e^{(t)}, s\right), \tag{3}$$

where the superscript (t) denotes the tth iteration. Eqs. (2) and (3) embody the idea of "iterative refinement" and have been widely used to recover e and s in (Lin and Tang 2007; Zhou et al. 2009; Li et al. 2013).

Due to its simplicity in model building and convenience in calculation, we continue to adopt the joint PDF $p(e, s)$ as a guide to explore the problem of face recognition with occlusion.

3 Markov random field for occlusion support

We first consider the priori probabilistic mass function (PMF) $p(s)$. For the binary random variable $s_i \in \{-1, 1\}$, its probability distribution can be naturally written as Bernoulli distribution

$$p(s_i) = \text{Bern}(s_i|\mu) = \mu^{\ddot{s}_i}(1-\mu)^{\dot{s}_i},$$

where $0 \le \mu \le 1$ is the probability with which pixel y_i is occluded. However, it makes no sense to consider the occlusion probability of an isolated pixel y_i, just as it is meaningless to consider the probability value of a continuous variable. The occlusion support s is actually a special random variable: numerically discrete but also spatially contiguous. (Lin and Tang 2007) were of the first to note this spatial contiguity. Spatial contiguity means that the occluded probability of a target pixel is determined by two aspects of its neighborhood pixels: the values of their occlusion supports and the distances from the target pixel. With the contiguous constraint and the independent identical distribution (i.i.d.) assumption, (Lin and Tang 2007) formulated $p(s)$ using a Markov Random Field (MRF)

$$p(s) \propto \prod_{i\in\mathcal{P}} \mu^{\ddot{s}_i}(1-\mu)^{\dot{s}_i} \prod_{j\in\mathcal{N}(i)} \exp\left(-\frac{d_{ij}^2}{\sigma_d^2}\left(1 - s_i s_j\right)\right), \tag{4}$$

where d_{ij} is the Euclidean distance between two pixels i and j, and $\mathcal{N}(i)$ is pixel i's neighborhood. As d_{ij} is almost constant for all $j \in \mathcal{N}(i)$, substitute $\frac{d_{ij}^2}{\sigma_d^2}$ with λ_s and (4) can be reduced to

$$p(s) \propto \exp\left(\sum_{i\in\mathcal{P}} \lambda_\mu s_i + \sum_{i\in\mathcal{P}} \sum_{j\in\mathcal{N}(i)} \lambda_s s_i s_j\right), \tag{5}$$

where $\lambda_\mu = \frac{1}{2}\log\frac{\mu}{1-\mu}$. Note that Eq. (5) is in accordance with the *Ising* model (Zhou et al. 2009), where $\sum_{i=1}^{m} \lambda_\mu s_i$ is the data energy and $\sum_{i=1}^{m}\sum_{j\in\mathcal{N}(i)} \lambda_s s_i s_j$ is the smooth energy.

Except for the contiguous structure, (Li et al. 2013) also incorporated the regular boundary structure into (5) and formulated a morphological graph model.

4 Conditional probabilistic error model in IGO domain

We now concentrate on the conditional PDF $p(e|s)$ in the IGO domain. In existing works, the conditional PDF $p(e|s)$ has been considered mainly in the original pixel domain with emphasis on the error distribution in the non-occluded region, where the errors are usually assumed to be small and centered on zero. However, in real world, the non-occluded region might also have huge variations, such as pose and illumination changes, and meanwhile the probabilistic error model in the occluded region is also very important to understand the generative mechanism of occlusion. The works of (Zhang et al. 2009) and (Tzimiropoulos et al. 2012) about the features of the image gradient orientations (IGO) provided a new solving route for the above two problems. In this section, we therefore mainly consider the conditional probability of the reconstruction error in the IGO domain, that is:

$$\begin{aligned} \widehat{e} &= \widehat{y} - \widehat{\hat{y}} \\ &= \widehat{y} - \widehat{A}x \\ &= \phi(y) - \phi(A)\cdot x \end{aligned} \tag{6}$$

where $\boldsymbol{\Phi}(\cdot)$ denotes the IGO transformation function, $\widehat{y}$ denotes the IGO feature of y, or just means its measuring unit is in radians, and $x \in \mathbb{R}^n$ is the coding coefficient of $\widehat{y}$ w.r.t. $\widehat{A}$.

4.1 Uniform error distribution in the occluded region

Image Gradient Orientations (IGO) were also called *Gradientfaces* in face recognition literature (Zhang et al. 2009). While (Zhang et al. 2009) illustrated that gradientfaces were robust to illumination changes by using the Lambertian reflectance model, (Tzimiropoulos et al. 2012) showed IGOs were also robust to occlusion and misalignment. Central to the methodology of (Tzimiropoulos et al. 2012) is the priori distribution of gradient orientation differences between two dissimilar images. We might intuitively think such a distribution is unpredictable and irregular. However, quite the opposite, (Tzimiropoulos et al. 2012) indicated that *the gradient orientation differences* $\widehat{d} \in \mathbb{R}^m$ *of any two pixel-wise dissimilar images* $a \in \mathbb{R}^m$

and $b \in \mathbb{R}^m$ *follow a uniform distribution in the interval* $[-\pi, \pi)$ *with a high significant level*. According to this statistical result, we can formulate that the error distribution in the occluded region ($s_i = 1$) as follows

$$\ddot{p}\left(\widehat{e}_i\right) = p\left(\widehat{e}_i | s_i = 1\right) = \frac{1}{2\pi}. \tag{7}$$

In spite of its simplicity and experimental reliability, the conditional PDF (7) is actually built on two conditions: *i)* the priori model of the occlusion support s should be solid and exact enough, and *ii)* the two comparative images should be point-wise dissimilar in the occluded region, which requires the reconstruction image should be accurate enough. We have explored $p(s)$ in Sect. 3 and will discuss the coding scheme of y w.r.t. A in Sect. 5.1. Here, it is interesting to explore how the uniform distribution (7) was used by Tzimiropoulos *et al.* to approximately cancel out the error caused by occlusion without knowing the occlusion support. Their method was built on the following theorem (Tzimiropoulos et al. 2012):

Theorem 1 Let $u(\cdot)$ be a mean ergodic stochastic process and $u(t)$ follows a uniform distribution in $[-\pi, \pi)$, then for any non-empty interval $\mathcal{X} \in \mathbb{R}$, $\int_{\mathcal{X}} \cos(u(t))dt = 0$.

Let $u\left(\widehat{e}\right) = \widehat{e}$, (Tzimiropoulos et al. 2012) showed that the distance between $\widehat{y}$ and $\widehat{\hat{y}}$ could be measured by

$$\begin{aligned} d^2\left(\widehat{y}, \widehat{\hat{y}}\right) &\triangleq \sum_{k\in\mathcal{P}} \left(1 - \cos\left(\widehat{y}_k - \widehat{\hat{y}}_k\right)\right) \\ &= \sum_{k\in\mathcal{P}} \left(1 - \cos\,\widehat{e}_k\right) \\ &= \sum_{k\in\mathcal{P}} 1 - \sum_{k\in\dot{\mathcal{P}}} \cos\widehat{e}_k - \sum_{k\in\ddot{\mathcal{P}}} \cos\widehat{e}_k \\ &\approx |\mathcal{P}| - c\left|\dot{\mathcal{P}}\right| - \int_{-\pi}^{\pi} \cos(t)dt \\ &= |\mathcal{P}| - c\left|\dot{\mathcal{P}}\right| - 0, \end{aligned} \tag{8}$$

where $c \in [-1, 1]$, $|\mathcal{P}|$ is the cardinality of the set $\mathcal{P}$, and the error $\widehat{e}_{\dot{\mathcal{P}}}$ in the non-occluded region is assumed to be very small. By applying the cosine kernel, Eq. (8) seems to skillfully eliminate the effect caused by occlusion without knowing the occlusion support. Nevertheless, as shown in Fig. 11 of (Tzimiropoulos et al. 2012), with occlusion level increasing, the quality of the reconstruction image $\hat{y}$ decreases sharply and it becomes more and more difficult to eliminate the occlusion influence only using Eq. (8).

4.2 Weight-conditional gaussian error distribution in the non-occluded region

We now consider the PDF $\dot{p}\left(\widehat{e}_i\right) = p\left(\widehat{e}_i | s_i = -1\right)$ of the reconstruction error $\widehat{e}_{\dot{\mathcal{P}}}$ in the non-occluded region, where $\widehat{e}_{\dot{\mathcal{P}}}$ might be caused by Gaussian white noise, misalignment, expression variant, illumination change and etc. Unlike facial occlusion, these variations are irregular but not point-wise dissimilar. Therefore, the uniform distribution is not applicable any more.

To explore the distribution of the errors incurred by irregular outliers, (Yang et al. 2011) summarized the common characteristics of the error distribution and proposed a generalized error distribution model, which finally lead to the so-called Robust Sparse Coding (RSC). Meanwhile, (He et al. 2011) studied the robust error metrics and proposed a CorrEntropy-based Sparse Representation (CESR) for robust face recognition. Despite starting from different viewpoints, both RSC and CESR are finally reduced to a weighted least squares (WLS) problem. Without considering the error coding schemes, the WLS problem can be summarized as

$$\min_{e,w} \sum_{i\in\mathcal{P}} \left(\left(w_i e_i\right)^2 + \lambda_w \omega\left(w_i\right)\right), \tag{9}$$

where $\lambda_w \geq 0$ is a regularized constant, $w_i \geq 0$ is a weight variable, and $\omega\left(w_i\right)$ is a cost function imposed on w_i. In (Yang et al. 2011), $\omega\left(w_i\right)$ was not explicitly formulated but w_i was intuitively set to a logistic function $w_i = \frac{1}{1+\exp\left(\mu e_i^2 - \mu\delta\right)}$, where $\mu > 0$ and $\delta > 0$ are two controlling parameters. In (He et al. 2011), $\omega\left(w_i\right)$ was set to a convex conjugate function of the Gaussian function

$$g(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right), \tag{10}$$

where $\sigma > 0$ is the Gaussian kernel size, which induced $w_i = \sqrt{g\left(e_i\right)}$.

The WLS formulation (9) inspires us to build the PDF of the reconstruction error $\widehat{e}_i$ in the IGO domain as a Gaussian distribution conditioned on the a weight w:

$$\dot{p}\left(\widehat{e}_i \middle| w_i\right) = p\left(\widehat{e}_i \middle| w_i, s_i = -1\right) \propto g\left(w_i \widehat{e}_i\right), \tag{11}$$

$$p\left(w_i\right) \propto \exp\left(-\lambda_w \omega\left(w_i\right)\right), \tag{12}$$

where we set $\omega\left(w_i\right)$ to be the convex conjugate function of the Gaussian function. As the weight w is independent of the occlusion support s, we have

$$\begin{aligned}\dot{p}\left(\widehat{e}_i, w_i\right) &= \dot{p}\left(\widehat{e}_i \middle| w_i\right) p\left(w_i\right), \\ &\propto g\left(w_i \widehat{e}_i\right) \exp\left(-\lambda_w \omega\left(w_i\right)\right) \\ &\propto \exp\left(-\frac{\left(w_i \widehat{e}_i\right)^2}{2\sigma^2} - \lambda_w \omega\left(w_i\right)\right).\end{aligned} \tag{13}$$

In the experimental section, we will demonstrate that the weight tuning is very critical for performance enhancing.

4.3 The holistic conditional error distribution

Substituting (7) and (11) into (1), we have the following holistic conditional PDF

$$\begin{aligned}p\left(\widehat{e}, w \middle| s\right) &= p\left(\widehat{e}_{\dot{\mathcal{P}}}, \widehat{e}_{\ddot{\mathcal{P}}}, w \middle| s\right) \\ &= \prod_{i\in\mathcal{P}} p\left(\widehat{e}_i \middle| s_i = 1\right) p\left(\widehat{e}_i, w_i \middle| s_i = -1\right) \\ &\propto \prod_{i\in\mathcal{P}} \left(\frac{1}{2\pi}\right)^{\ddot{s}_i} \exp\left(-\frac{\dot{s}_i\left(w_i \widehat{e}_i\right)^2}{2\sigma^2} - \lambda_w \dot{s}_i \omega\left(w_i\right)\right) \\ &\propto \exp\left(-\sum_{i\in\mathcal{P}} \frac{1}{2}\left(1 - s_i\right)\left(\left(w_i \widehat{e}_i\right)^2 + \tilde{\lambda}_w \omega\left(w_i\right)\right)\right. \\ &\left. -\tilde{\lambda}_\sigma \sum_{i\in\mathcal{P}} s_i\right),\end{aligned} \tag{14}$$

where $w = \left[w_{i\in\mathcal{P}}\right]$ is the weight vector, $\tilde{\lambda}_w = 2\sigma^2\lambda_w$, and $\tilde{\lambda}_\sigma = \sigma^2 \log 2\pi$.

Note that the energy term (i.e., the negative of the exponential part) of (14) is actually an extension of the distance defined in Eq. (8) (see also Eq. (6) of (Tzimiropoulos et al. 2012)). To see this, we expand (8) as follows

$$\begin{aligned}\sum_{i\in\mathcal{P}} \left(1 - \cos \widehat{e}_i\right) &= \sum_{i\in\mathcal{P}} \dot{s}_i \left(1 - \cos \widehat{e}_i\right). \\ &+ \sum_{i\in\mathcal{P}} \ddot{s}_i \left(1 - \cos \widehat{e}_i\right)\end{aligned} \tag{15}$$

$$\approx \sum_{i\in\mathcal{P}} \frac{1}{2} \dot{s}_i \widehat{e}_i^{\,2} + \sum_{i\in\mathcal{P}} \ddot{s}_i - \sum_{i\in\mathcal{P}} \ddot{s}_i \cos \widehat{e}_i \tag{16}$$

$$= \frac{1}{2}\left(\sum_{i\in\mathcal{P}} \frac{1}{2}\left(1 - s_i\right)\widehat{e}_i^{\,2} + \sum_{i\in\mathcal{P}} s_i + |\mathcal{P}|\right). \tag{17}$$

From (15) to (16), we use the equivalent infinitesimal of the cosine function (i.e., $1 - \cos x \sim \frac{1}{2}x^2$), as the errors in the non-occluded region is assumed to approximate zero. Both (14) and (17) say that for the known occlusion support s, the total energy of the reconstruction error $\widehat{e}$ depends only on the non-occluded region. It further explains why the existing works put emphasis on the model of the reconstruction error in the non-occluded region.

5 IGO-embedded structural error coding model

5.1 The proposed model

By combining (14) and (5), we have the following joint generative model

$$\begin{aligned}p\left(\widehat{e}, w, s\right) &\propto p\left(\widehat{e}, w \middle| s\right) p(s), \\ &\propto \exp\left(-\sum_{i\in\mathcal{P}} \frac{1}{2}\left(1 - s_i\right)\left(\left(w_i \widehat{e}_i\right)^2 + \tilde{\lambda}_w \omega\left(w_i\right)\right) + \right.\end{aligned} \tag{18}$$

$$\left. \sum_{i\in\mathcal{P}} \lambda_d s_i + \sum_{i\in\mathcal{P}} \sum_{j\in\mathcal{N}(i)} \lambda_s s_i s_j\right) \tag{19}$$

where $\lambda_d = \lambda_\mu - \tilde{\lambda}_\sigma$.

In order to estimate the optimal $\widehat{e}$, w and s from the generative model, we use the criterion of maximizing the joint PDF $p\left(\widehat{e}, w, s\right)$. According to (6), the reconstruction error $\widehat{e}$ is actually calculated by $\widehat{y}$ and its linear reconstruction $\widehat{\hat{y}}$ w.r.t. the training set $\widehat{A}$. Hence, the linear coding scheme $\widehat{Ax}$ is very important to determine the reconstruction quality of $\widehat{y}$. As sparse coding has been shown to be robust to outliers (Wright et al. 2009; He et al. 2011), we adopt sparse representation framework to calculate x. Generally, there are two convex constraints leading to sparse coding: the ℓ_1-norm minimization (Wright et al. 2009; Yang et al. 2011) and the nonnegative constraint (He et al. 2011). We here just choose the nonnegative sparse constraint for efficient calculation. Now, we summarize the proposed error coding model as follows

$$\begin{aligned}\left(\widehat{e}^{\,*}, w^*, s^*\right) &= \arg\max_{\widehat{e}, s} p\left(\widehat{e}, w, s\right) \\ s.t. \quad \widehat{e} &= \widehat{y} - \widehat{A}x, x \geq 0.\end{aligned} \tag{20}$$

Since the above error coding model combines the IGO features and the two weights w and s to describe the error structure, we refer the optimization problem (20) as an *IGO-embedded Structural Error Coding* (IGO-SEC) model.

5.2 Iterative algorithm

Since there are 3 variables $\widehat{e}$, w and s in the objective function (20) of the IGO-SEC model, it is difficult to optimize them simultaneously. We therefore adopt the alternating maximization way, as shown by (2) and (3), to solve a local maximizer $\left(w, \widehat{e}, s\right)$ of (20)

$$w^{(t)} = \arg\max_{w} -\sum_{i\in\mathcal{P}} \frac{1}{2}\left(1 - s_i^{(t-1)}\right)\left(\left(w_i \widehat{e}_i^{(t-1)}\right)^2 + \tilde{\lambda}_w \omega\left(w_i\right)\right), \tag{21}$$

$$\widehat{e}^{(t)} = \arg\max_{\widehat{e}} -\sum_{i\in\mathcal{P}} \frac{1}{2}\left(1 - s_i^{(t-1)}\right)\left(w_i^{(t)} \widehat{e}_i\right)^2 \quad s.t.\ \widehat{e} = \widehat{y} - \widehat{A}x, x \geq 0, \tag{22}$$

$$s^{(t)} = \arg\max_{s} \sum_{i\in\mathcal{P}} s_i \left(\lambda_d + \frac{1}{2}\left(w_i^{(t)} \widehat{e}_i^{(t)}\right)^2 + \frac{1}{2}\tilde{\lambda}_w \omega\left(w_i^{(t)}\right)\right) + \sum_{i\in\mathcal{P}} \sum_{j\in\mathcal{N}(i)} \lambda_s s_i s_j. \tag{23}$$

Eq. (23) is in accord with the classical Ising model and can be solved using graph cuts (Kolmogorov and Zabih 2004). We mainly concern the optimization problem (21) and (22).

To solve (21), we note that the weight vector w is a hidden variable for the generative error model (18) and stems from the maximum correntropy criterion proposed in (He et al. 2011), where the half-quadratic technique is used to solve w. According to this technique, we could directly obtain w as follows

$$w_i^{(t)} = g\left(\frac{1}{2}\left(1 - s_i^{(t-1)}\right)\widehat{e}_i^{(t-1)}\right). \tag{24}$$

To solve (22), we first rewrite (22) as an optimization problem about x in the form of matrix manipulation

$$x^{(t)} = \arg\min_{\widehat{e}} \left\| W^{(t)} \dot{S}^{(t-1)} \left(\widehat{y} - \widehat{A}x\right)\right\|_2^2, s.t.\ x \geq 0 \tag{25}$$

where $\dot{S}^{(t-1)} = \mathrm{diag}\left(\frac{1-s^{(t-1)}}{2}\right)$ and $W^{(t)} = \mathrm{diag}\left(w^{(t)}\right)$. Here $\mathrm{diag}(\cdot)$ is an operator to convert a vector to a diagonal matrix. For the known $W^{(t)}$ and $\widetilde{S}^{(t-1)}$, (25) is a nonnegative least squares problem (NNLS) and can be solved by the classical active set algorithm (Portugal et al. 1994).

6 Two ways to improve the error image in IGO-SEC

Clearly, among the three variables $\widehat{e}, w, s$ of IGO-SEC, the error image $\widehat{e}$ plays a central role. However, in Eq. (22), $\widehat{e}$ is simply calculated as the difference between $\widehat{y}$ and its reconstruction $\widehat{y} = \widehat{A}x$. This calculation does not fully consider the effect of the IGO features. In this section, we suggest two methods to improve the error image in IGO by building a high-quality reconstruction image $\widehat{y}$ and fully using the structure of the IGO features.

6.1 Reconstruction image boosting: from Pixel to IGO

Given a coding coefficient x, we actually have two distinct ways to calculate the reconstruction image $\widehat{y}$ in the IGO domain:

$$\widehat{y}^{1} = \boldsymbol{\Phi}(A)x \tag{26}$$

$$\widehat{y}^{2} = \boldsymbol{\Phi}(Ax) \tag{27}$$

Sepecifically, Eq. (26) directly calculates $\widehat{y}$ in the IGO domain, whereas Eq. (27) first calculates the reconstruction image in the pixel domain and then transforms it into the IGO domain.

Although there is only a little difference between Eqs. (26) and (27), their calculating results $\widehat{y}^{1}$ and $\widehat{y}^{2}$ are very different, as shown in Fig. 1a. It seems that $\widehat{y}^{1}$ contains many tiny noises and thus is visually dissimilar with the ground truth $\widehat{y}^{0}$, while $\widehat{y}^{2}$ is much smoother and seems more similar to $\widehat{y}^{0}$. We further measure the distances between $\widehat{y}^{1}/\widehat{y}^{2}$ and $\widehat{y}^{0}/\widehat{y}$ by ℓ_2 and d^2 in Fig. 1b. Interestingly, ℓ_2 shows that $\widehat{y}^{1}$ is closer to $\widehat{y}^{0}/\widehat{y}$ than $\widehat{y}^{2}$, while d^2 gives contrary results. Clearly, the results of d^2 is in accordance with human vision. To further compare $\widehat{y}^{1}$ and $\widehat{y}^{2}$, Fig. 1c plots the error images $\widehat{e}^{1} = \widehat{y}^{0} - \widehat{y}^{1}$ and $\widehat{e}^{2} = \widehat{y}^{0} - \widehat{y}^{2}$ and their post-processed results by two structural error metrics CEM and CC, which will be introduced in the next subsection. As can be seen, $\widehat{e}^{2}$ and its post-processed results have much better local clustering effects (i.e., better to separate occlusion from non-occlusion) than the ones about $\widehat{e}^{1}$.

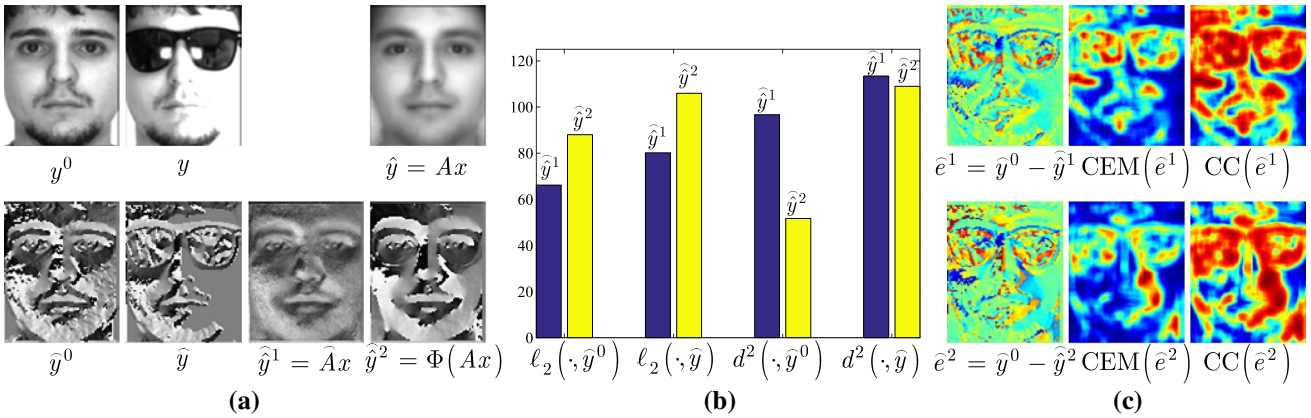


Fig. 1 **a** Visual comparison of the ground truth image $y^0(\widehat{y}^0)$, the test image y $(\widehat{y})$ with sunglasses disguise, and its reconstruction images $\widehat{\hat{y}}^1$ and $\widehat{\hat{y}}^2$ calculated by (26) and (27), respectively. **b** Numerical comparison of the distances from $\widehat{\hat{y}}^1$ /$\widehat{\hat{y}}^2$ to $\widehat{y}^0$ and $\widehat{y}$, computed by ℓ_2 (i.e., Euclidean distance) and d^2 (see Eq. (8)), respectively. **c** Visual comparison of the error images $\widehat{e}^1$and $\widehat{e}^2$using the error metrics $\mathcal{E}_s$ and $\mathcal{E}_c$

Then, why $\widehat{\hat{y}}^2$ is better than $\widehat{\hat{y}}^1$? The main reason lies in the numerical characteristic of the IGO features. The IGO feature is gained by calculating the relative difference between vertical and horizontal pixels in each neighborhood. This means that an image in IGO domain usually has a bigger total variation (TV) than its original form. Therefore, for a given coding coefficient x, the reconstruction image in IGO , i.e., $\boldsymbol{\Phi}(\mathrm{A})x$, usually has a bigger TV than Ax. Due to superimposing less variations, $\boldsymbol{\Phi}(\mathrm{Ax})$ tends to be more stable and much smoother than $\boldsymbol{\Phi}(\mathrm{A})x$. Experiments in Sect. 8 will see the great benefits brought by $\boldsymbol{\Phi}(\mathrm{Ax})$ over $\boldsymbol{\Phi}(\mathrm{A})x$.

6.2 Error image boosting by using structural error metrics

To further use the robust features embedded in IGO, we now explore how to post-process the error image $\widehat{e} = \widehat{y} - \widehat{\hat{y}}$ by using the existing robust error/distance metrics. Note that the output of an error metric is not a matrix but a scalar. To use an error metric to reformulate the error image, we suggest the following scheme: firstly, regarding each pixel and its closed neighbors in the error image as a macro pixel or a sub-image[1], and then applying the error metric to each macro pixel to get a new error value[2], and finally combining these new error values into a new error image. As per this scheme, what is critical to form a high-quality error image is to seek for a suitable robust error metric.

As indicated in (He et al. 2011), a robust error metric is local, that is: it highlights the dissimilarities of the two compared images and simultaneously weakens their similarities. Based on this idea, many robust error metrics, such as M-estimators (He et al. 2011, 2014), have been developed. Among them, the correntropy induced metric (CIM) (Liu et al. 2007; He et al. 2011; Wang and Wang 2013) is the most widely used one, which is defined as follows

$$\mathrm{CIM}(e) \triangleq 1 - \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} g(e_i). \tag{28}$$

Note that $g(\cdot)$ in CIM is a Gaussian kernel and CIM is usually used in the pixel domain. In the IGO domain, however, Theorem 1 and Eq. (8) indicate that applying the cosine kernel to the error image could lead to a robust distance measure. Thus, to form a robust error metric in IGO, we can shrink the action scope of Eq. (8) to the neighborhood of each pixel in the error image, that is, we adopt the following cosine-based error metric (CEM) to each macro pixel of the error image

$$\mathrm{CEM}\left(\widehat{e}_i\right) \triangleq \sum_{j\in\mathcal{N}(i)} \left(1 - \cos\left(\widehat{e}_j\right)\right), \tag{29}$$

and then organize them to form a new error image. Actually, the new error image can be fast computed by applying mean filtering to the transformed domain $f = 1 - \cos\left(\widehat{e}\right)$:

$$\mathrm{CEM}\left(\widehat{e}\right) \triangleq \left(1 - \cos\left(\widehat{e}\right)\right) * \mathbf{1}_{|\mathcal{N}|\times|\mathcal{N}|}, \tag{30}$$

where * denotes a convolutional operator and $\mathbf{1}_{|\mathcal{N}|\times|\mathcal{N}|}$ denotes an $|\mathcal{N}| \times |\mathcal{N}|$ matrix with all-1 elements.

It's interesting to compare CEM and CIM. In fact, CEM harnesses the spatial local information of the image

[1] Such a processing method is very common in image processing (Ahonen et al. 2006; Qian et al. 2013).

[2] Note that it makes nonsense to impose a robust error metric on each pixel.

Table 1 Eight variants of IGO-SEC

Variants	1	2	3	4	5	6	7	8
Choices	$\widehat{y}^{2}$ +CC+w	$\widehat{y}^{2}$ +CC	$\widehat{y}^{2}$ +CEM+w	$\widehat{y}^{2}$ +CEM	$\widehat{y}^{1}$ +CC+w	$\widehat{y}^{1}$ +CC	$\widehat{y}^{1}$ +CEM+w	$\widehat{y}^{1}$ +CEM

difference, while CIM integrates the statistical local information. We therefore further feed the result of CEM to CIM and formulate the final error image

$$\mathrm{CC}\left(\widehat{e}\right) \triangleq \mathrm{CIM}\left(\mathrm{CEM}\left(\widehat{e}\right)\right), \tag{31}$$

which simultaneously integrates the spatial and statistical local information. Since Eq. (30) and (31) finally lead to an error image but not a scalar, we refer to them as *structural error metrics* to distinguish from the common error metrics. Fig. 1c shows that CC outperforms CEM in source-and-error separation.

7 Final algorithm of IGO-SEC

Now we have several ways to solve IGO-SEC. Basically, we can iteratively alternate equations (21), (22) and (23) to obtain the final solution. But we can also replace the reconstruction image $\widehat{y}$ in (22) by $\hat{y}^{1}$ or $\hat{y}^{2}$, post-process the error image $\widehat{e}$ by CEM or CC, use or not use the weight w. Each choice will lead to a new variant of IGO-SEC. The total 8 variants are listed in Table 1. The experiments in Sect. 8 showed that reconstruction method $\hat{y}^{2}$ generally outperformed $\hat{y}^{1}$. We therefore mainly consider the first four $\hat{y}^{2}$-based IGO-SECs in Table 1: IGO-SEC-1, IGO-SEC-2, IGO-SEC-3, and IGO-SEC-4. For the $\hat{y}^{1}$-based variants, we only consider the optimal one, IGO-SEC-5.

Algorithm 1 Algorithm of IGO-SEC

Input: training data A, test sample y, total iteration number T, reconstruction operator $\mathcal{R}(\cdot,\cdot)$, structural error metric $\mathcal{E}(\cdot)$
// *Here,* $\mathcal{R}(\cdot,\cdot)=\Phi(A)\,x$ *or* $\mathcal{R}(\cdot,\cdot)=\Phi(Ax)$, *see Subsection 6.1*
// *Here,* $\mathcal{E}(\cdot)=CEM(\cdot)$ *or* $\mathcal{E}(\cdot)=CC(\cdot)$, *see Subsection 6.2*

Output: estimated error $\widehat{e}$, estimated occlusion support s, identity (y)

1. Calculate the mean face $\bar{y}$ of the training image A;
2. Transform y, $\bar{y}$ and A into the IGO domain: $\widehat{y}=\Phi(y)$, $\widehat{A}=\Phi(A)$;
3. Initialize the coding coefficient $x^{(0)}=\frac{1}{n}\mathbf{1}_n$;
4. Initialize the reconstruction error $\widehat{e}^{(0)}=\mathcal{E}\left(\widehat{y}-\mathcal{R}\left(A,x^{(0)}\right)\right)$;
5. Initialize the occlusion support $s^{(0)}=\mathcal{K}\left(\widehat{e}^{(0)}\right)$;
6. **For** $t=1$ **to** T
7. Calculate the weight $w^{(t)}$, where $\forall i\in\mathcal{P}$, $w_i^{(t)}=g\left(\frac{1}{2}\left(1-s_i^{(t-1)}\right)\widehat{e}_i^{(t-1)}\right)$;
8. Calculate the coding coefficient $x^{(t)}$ of $\widehat{y}$ w.r.t. $\widehat{A}$ according to (25) using the active set algorithm;
9. Calculate the reconstruction error $\widehat{e}^{(t)}=\mathcal{E}\left(\widehat{y}-\mathcal{R}\left(A,x^{(t)}\right)\right)$;
10. Recover the occlusion support $s^{(t)}$ according to (23) by GraphCuts;
11. **End For**
12. Set $x=x^{(T)}$, $s=s^{(T)}$, $\dot{S}=\mathrm{diag}\left(\frac{1-s}{2}\right)$, $W=\mathrm{diag}\left(w^{(T)}\right)$, $\widehat{e}=\widehat{y}-\widehat{A}x$;
13. For $k=1,\cdots,K$, compute the residuals $r_k=\left\|W\cdot\dot{S}\cdot\left(\widehat{y}-\widehat{A}\delta_k(x)\right)\right\|_2^2$, where $\delta_k(x)$ is a new vector whose only nonzero entries are the ones in x that correspond to subject k;
14. identity $(y)=\arg\min_k r_k$.

Before getting the final algorithm of 1, we still need to handle two detailed problems. First, how to initialize IGO-SEC. Although there are 3 unknown variables $\widehat{e}, w, s$ in IGO-SEC, we only need to initialize a hidden variable x. We adopt the strategy proposed in (Yang et al. 2011), where $x^{(0)}$ is set to $\frac{1}{n}\mathbf{1}_n$ (i.e., the coding scores are uniformly assigned to each atom the dictionary A). Another detail is how to classify the test image from existing training subjects by using the final solution of IGO-SEC. We adopt a subject specific reconstruction classifier similar to the sparse representation-based classifier proposed by (Wright et al. 2009), with the major difference that the classifier introduced here is based on the IGO-SEC model. The final implementation of IGO-SEC is provided in Algorithm 1.

8 Simulations and experiments

To evaluate the performance of the proposed IGO-SEC model, we compare it with 5 robust classifiers (CRC (Zhang et al. 2011), CESR (He et al. 2011), RSC (Yang et al. 2011), SSEC (Li et al. 2013) and NMR (Yang et al. 2017)) and 3 robust deep learning methods (VGG (Parkhi et al. 2015), LCNN (Wu et al. 2015), and PCANet(Chan et al. 2015)). Experiments were performed on four public face databases, namely, the Extended Yale B (Georghiades et al. 2001), AR (Martinez 1998), UMB-DB (Colombo et al. 2011), and LFW (Huang et al. 2007) databases. The recognition rates of the compared methods will be reported and analyzed. All algorithms, except for VGG and LCNN, were implemented in MATLAB on an Intel Dual-Core 2.80 GHz Windows 10 machine with 8 GB memory. VGG and LCNN were tested in MatCaffe by calling the Caffe models provided by their authors[3] on an Intel 8-Core 2.4 GHz Ubuntu 16.04.2 with 64 GB memory and 2 Tesla K40c (12GB) GPUs.

8.1 Experimental setting and face databases

Algorithm Setting. Among many existing face recognition CNN models, we choose VGG and LCNN, because the two CNN models are freely available and their network structures are similar to DeepID (Taigman et al. 2015; Sun et al. 2016)[4] and DeepFace (Taigman et al. 2015), which were claimed to be robust to facial occlusions to some extent. As VGG and LCNN are mainly used to extract facial features, we choose CRC (Zhang et al. 2011) as their classifier. For PCANet, we select nearest neighbor (NN) equipped with Chi-square distance as its classifier, as suggested by the authors (Chan et al. 2015). For fair comparison, we use the IGO features rather than the pixel features in the 5 robust classifiers. Specifically, we will compare IGO-SEC with the following 8 methods: CRC+IGO, CESR+IGO, RSC+IGO, SSEC+IGO, NMR+IGO, CRC+VGG, CRC+LCNN and NN+PCANet. The parameters of all compared methods are set according to the strategy suggested in their papers.

It is worth noticing how to choose the parameters of IGO-SEC. As seen from the main iterative equations (21), (22) and (23), there are 3 main parameters: $\tilde{\lambda}_w$, λ_d and λ_s. $\tilde{\lambda}_w$ is used to regularize the weight w. As w can be directly solved by the half-quadratic technique (see Eq. (24)), we can neglect the value of $\tilde{\lambda}_w$. The two parameters λ_d and λ_s are both from the Ising model, where they are called data parameter and smooth parameter, respectively. In this paper, the data parameter λ_d sets the tendency of each pixel to be occluded or not. Since such a tendency is usually indicated by the pixel values of the error image, λ_d is usually set to 0. The smooth parameter λ_s sets the priority of each pixel to be occluded or not as per the labels of its neighboring pixels. Clearly, a too large λ_s might lead to all pixels are labeled as occlusion or non-occlusion, whereas a too small λ_s might totally neglect the spatial contiguity of occlusion. Thus, λ_s is usually set to 2. In addition, there are 3 secondary parameters in IGO-SEC: the two neighbor sets $\mathcal{N}$'s in Eqs. (23) and (29), and the total iteration number T in Algorithm 1. According to the experimental verification, $\mathcal{N}$ in Eq. (23) and (29) can be set to a 3×3 and 5×5 rectangular box around its central pixel, and $T = 6$ is enough for the convergence of IGO-SEC.

Database. We conduct a set of large-scale experiments on 4 benchmark databases:

1. Extended Yale B database (Georghiades et al. 2001): The Extended Yale B database consists of 2,414 frontal face images from 38 subjects (Georghiades et al. 2001) under various lighting conditions. The cropped and normalized 19,2168 face images were captured under 5 controlled lighting conditions (Lee et al. 2005): from normal illumination conditions to extreme ones.
2. AR database (Martinez 1998): The AR database consists of over 4,000 facial images from 126 subjects (70 men and 56 women). For each subject, 26 facial images are taken in two sessions (separated by two weeks). These images suffer different facial variations, including various facial expressions (neutral, smile, anger, and scream), illumination variations (left light on, right light on, and all side lights on), and occlusions (sunglasses and scarves). This database is always used for evaluating robust face recognition algorithm.

[3] The VGG and LCNN model can be downloaded from http://www.robots.ox.ac.uk/~vgg/software/vgg_face/ and https://github.com/AlfredXiangWu/face_verification_experiment, respectively.

[4] Note that the lastest version (Sun et al. 2016) of DeepID was totally based on VGG.

Fig. 2 Examples of the training and test images of Subject 1 from the Extended Yale B database. Row 1: clean training samples from Subset I and II; Rows 2 to 4: samples from Test Set 1–3 with 0–90% contiguous occlusion under various illumination condition Subset III, IV, and V

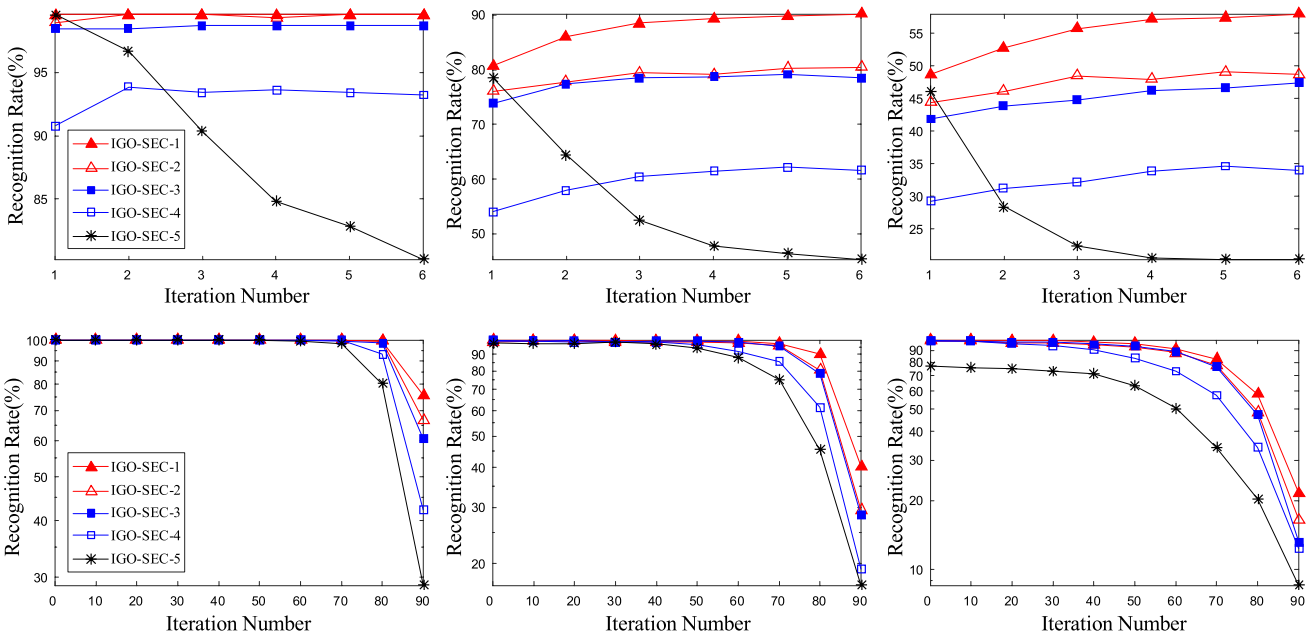


Fig. 3 Recognition results of 5 variants of IGO-SEC (i.e., IGO-SEC-i, $i \in \{1, 2, 3, 4, 5\}$) on three test sets (from let to right are respectively Test Set 1, 2 and 3) of the Extended Yale B database. Top Row: recognition rates versus 6 iterations against the 80% occlusion; bottom row: recognition rates at the 6th iteration versus various percents (0%~90%) of occlusions

3. UMB-DB database(Colombo et al. 2011): The UMB-DB database was originally built to test algorithms and systems for 3D face analysis in uncontrolled and challenging scenarios, particularly in those cases where faces are occluded. The database is composed of 1,473 pairs of depth and color images of 143 subjects. Each subject has been acquired with different facial expressions, and with the face partially occluded by various objects, such as eyeglasses, hats, scarves and hands. In

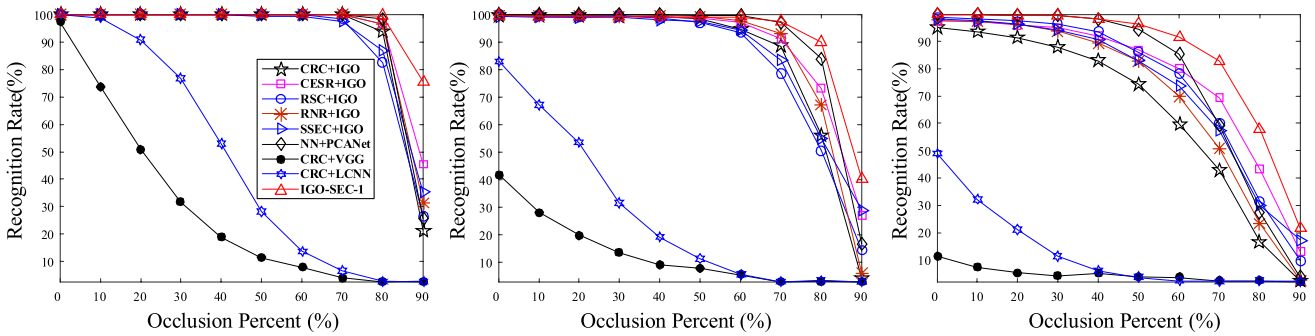


Fig. 4 Recognition results on three test sets (from let to right are respectively Test Set 1, 2 and 3) of the Extended Yale B database

Table 2 Recognition rates (%) with 60% to 90% occlusion percents (OP) on three test sets of the Extended Yale B

	Test set 1				Test set 2				Test set 3			
OP	60%	70%	80%	90%	60%	70%	80%	90%	60%	70%	80%	90%
CRC+IGO	100.00	100.00	93.85	21.32	94.87	88.59	56.27	4.18	59.80	42.72	16.67	2.38
CESR+IGO	100.00	100.00	96.48	45.49	97.34	91.44	73.19	27.38	79.97	69.33	43.14	13.03
RSC+IGO	100.00	98.46	82.64	26.59	93.54	78.71	50.76	14.45	78.15	60.08	31.37	9.66
NMR+IGO	100.00	100.00	96.70	31.21	97.72	93.16	67.11	6.08	70.03	50.70	23.25	2.66
SSEC+IGO	99.56	97.58	86.81	35.16	94.30	83.46	54.94	29.09	73.81	57.42	29.83	17.23
NN+PCANet	100.00	100.00	98.46	25.49	99.81	97.34	84.03	16.73	85.15	59.24	27.31	3.64
CRC+VGG	7.91	3.96	2.42	2.86	5.51	3.04	3.23	3.04	3.64	2.38	2.38	1.82
CRC+LCNN	13.85	6.59	2.86	2.42	5.89	3.04	3.61	2.85	2.38	2.10	2.38	2.24
IGO-SEC-1	100.00	100.00	99.56	75.60	99.24	97.46	90.08	40.27	91.52	82.67	57.86	21.53

this work, we only use the color images for 2D face recognition testing.

4. LFW database (Huang et al. 2007): The LFW database contains 13,233 face images of 5749 subjects collected from internet and has been widely used for unconstrained face recognition with variations of pose, illumination, expression, misalignment and occlusion, and so on.

8.2 Synthetic occlusions using the extended yale B database

In this experiment, we use the Extended Yale B database to test the robustness of our algorithm against various levels of synthetic occlusions mixed with various illuminations. The evaluation protocols are similar to (Wright et al. 2009; Yang et al. 2017; Li et al. 2017) and identical to (Wei et al. 2012). For training, we use the clean images from Subset I and II (717 images, with normal-to-moderate illumination conditions). For testing, we construct 3 test sets (namely, Test Set 1, Test Set 2 and Test Set 3), which consist of images from Subset III (453 images, with extreme illumination conditions), Subset IV (524 images, with more extreme illumination conditions) and Subset V (712 images, with the most extreme illumination conditions), respectively. In order to simulate different levels (from 0% to 90% with step size 10%) of contiguous occlusion, we also replace a randomly located square patch from each test image with a "baboon" image which has similar structure with the human face and has been widely used in robust face recognition testing (Wright et al. 2009; Zhou et al. 2009; Yang et al. 2011; He et al. 2011; Li et al. 2013, 2017; Yang et al. 2017). Thus, the sizes of the 3 test sets are enlarged by 10 folds. Note that the images in the training set and the test sets are non-overlapping. Figure 2 shows some examples of the training and test sets. All images are cropped and resized to 96×84 pixels.

We first compare the recognition performances of the 5 IGO-SECs versus various iterations and occlusion percents on the 3 test sets. Figure 3 plots the results. As seen from the top row, the two IGO-SECs with weight tuning (i.e., IGO-SEC-1 and IGO-SEC-3) have better convergences than those without weights tuning (i.e., IGO-SEC-2 and IGO-SEC-4). Such a superiority becomes more obvious with the recognition difficulty increasing. This demonstrates weight tuning is very critical for performance enhancing. On the other hand, we can also see that the reconstruction method $\hat{y}$ causes better performances than $\hat{y}$, and $\hat{y}$ even leads to performance dropping with

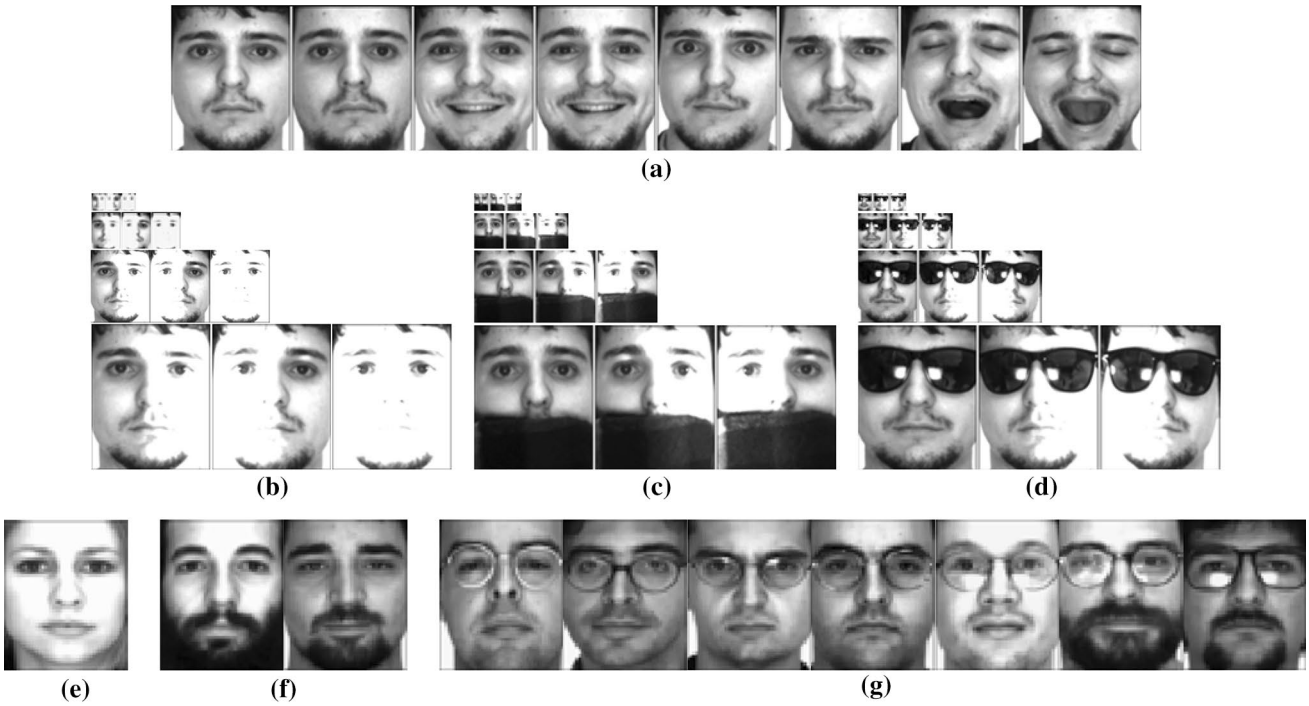


Fig. 5 Examples of the training and test sets of Subject 2 from the AR database: **a** the 8 clean training images from two sessions with normal illumination; **b**, **c**, **d** the test images simultaneously containing illumination variations / scarves disguises /sunglasses disguises under different feature dimensions (from top to bottom): 154, 644, 2,576, and 10,304

iteration increasing. The bottom row of Fig. 3 shows that with recognition difficulty increasing, the breakdown points of all compared methods become lower and lower. This shows that the illumination condition could greatly effect the accurate estimation of the occlusion support s. In sum, Fig. 3 points out that the recognition performances of the 5 IGO-SECs can be ranked as follows: IGO-SEC-1≥ IGO-SEC-2≥IGO-SEC-3≥IGO-SEC-4≥IGO-SEC-5. This ranking result also states that the structural error metric CC is more significant than CEM.

Then, we compare the recognition rates of the optimal IGO-SEC (i.e., IGO-SEC-1) with other compared approaches on the three test sets. Figure 4 plots the results. For more clarity, Table 2 further gives the detailed recognition rates against 60% to 90% occlusions. As can be seen, under various situations, IGO-SEC-1 gains the optimal recognition performance, and the performance advantages become more significant when the occlusion level is large or the illumination conditions become worse. By contrast, VGG and LCNN have much lower recognition rates than the other compared methods. The main reason is that VGG and LCNN were trained in unconstrained settings, where the data distribution has a big distinction with the one in constrained situations. In Sect. 8.5, we will show that the performance gains of VGG and LCNN over other methods are very obvious under unconstrained scenarios. (Ghazi and Ekenel 2016) and (Li et al. 2017) also gave similar demonstrations. Although PCANet also belongs to the deep learning methods, it does not depend on the large-scale training data and can be easily retrained with the new training set. Thus, PCANet performs much better than VGG and LCNN. CESR+IGO can be roughly considered as another variant of IGO-SEC without using the occlusion support information, and has the second highest recognition rates in most test cases. NMR+IGO also performs very well. This means that the nuclear norm regularization imposed on the reconstruction error plays a very important role in dealing with occlusion and other variations. However, due to not using any strategy to eliminate contiguous occlusion, NMR has a very big performance dropping when the occlusion level is very high.

8.3 Real-world occlusions using the AR database

In this experiment, we evaluate the robustness of our method in dealing with the real disguises on the AR database. The evaluation protocols are similar to (He et al. 2011; Yang et al. 2017; Li et al. 2017). We first select a subset of the dataset that consists of 119 subjects (65 males and 54 females) and then resize the grayscale images to resolution 112×92. For training, we select 952 non-occluded frontal view images with varying facial expressions but normal illuminations

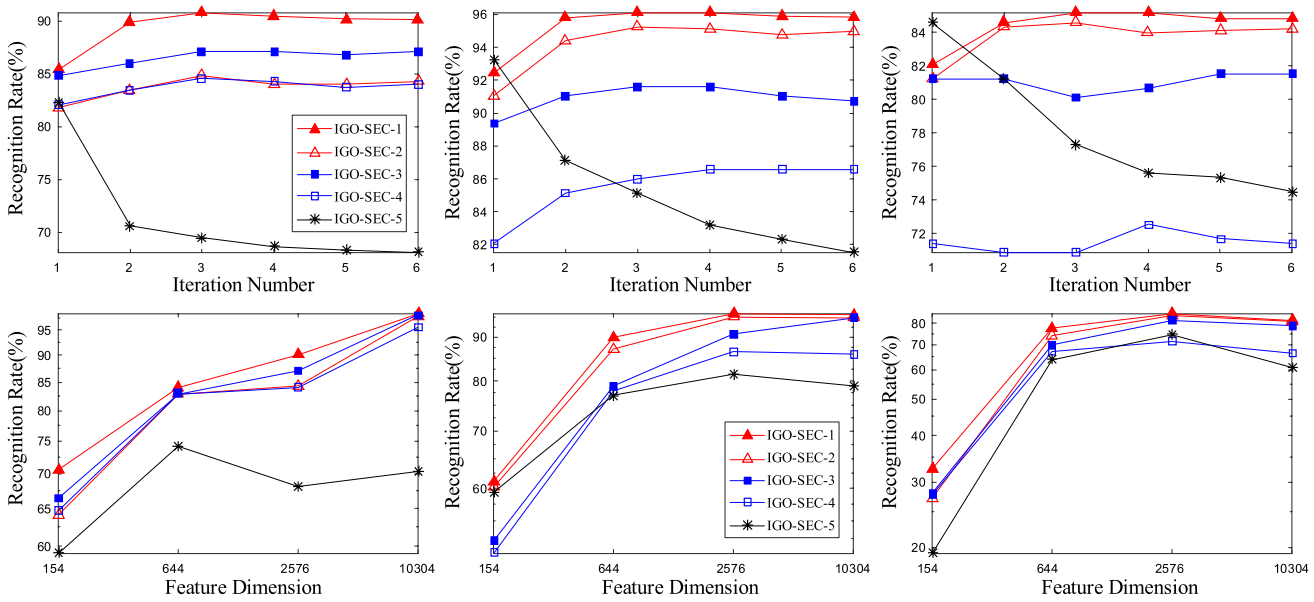


Fig. 6 Recognition results of 5 variants of IGO-SEC (i.e., IGO-SEC-i, $i \in \{1, 2, 3, 4, 5\}$) on three test sets (from let to right are respectively Test Set 1, 2 and 3) of the AR database. Top Row: recognition rates at the feature dimension 2,576 versus 6 iterations; bottom row: recognition rates at the 6th iteration versus various feature dimensions

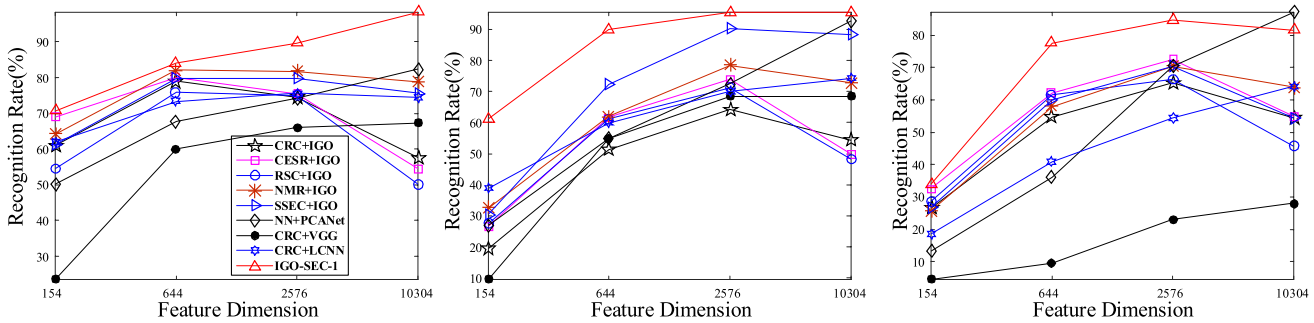


Fig. 7 Recognition results on three test sets (from let to right are respectively test set 1, 2 and 3) of the AR database

Table 3 Recognition rates (%) on three test sets of the AR database

	Test set 1				Test set 2				Test set 3			
FD	154	644	2576	10304	154	644	2576	10304	154	644	2576	10304
CRC+IGO	61.06	*78.99*	74.51	57.70	19.33	51.26	*64.15*	54.34	26.61	54.62	*65.27*	54.34
CESR+IGO	68.91	*80.11*	75.35	54.62	26.61	61.62	*73.95*	49.86	32.49	62.18	*72.55*	54.90
RSC+IGO	54.62	*75.91*	75.07	50.14	27.45	61.06	*71.15*	48.18	28.57	61.62	*66.39*	45.66
NMR+IGO	64.15	*82.07*	81.79	78.71	32.49	61.90	*78.43*	72.83	25.77	57.98	*70.31*	63.87
SSEC+IGO	61.34	79.83	*79.83*	75.63	30.53	72.27	*90.20*	88.24	26.89	59.94	*70.31*	54.34
NN+PCANet	50.14	67.79	74.23	*82.35*	26.89	54.90	72.55	*92.72*	13.17	36.13	70.31	***87.11***
CRC+VGG	23.53	59.94	66.11	*67.23*	9.52	54.90	68.35	*68.35*	4.48	9.52	22.97	*28.01*
CRC+LCNN	61.90	73.39	*75.63*	74.51	38.94	59.94	70.03	*74.23*	18.49	40.90	54.34	*64.15*
IGO-SEC-1	**70.59**	**84.03**	**89.64**	***98.32***	**61.06**	**89.92**	**95.52**	***95.52***	**33.89**	**77.59**	***84.59***	81.51

The maximum recognition rate of each method on various *feature dimensions* (FD) is marked by underlined italic fonts. The best result in each column is in bold

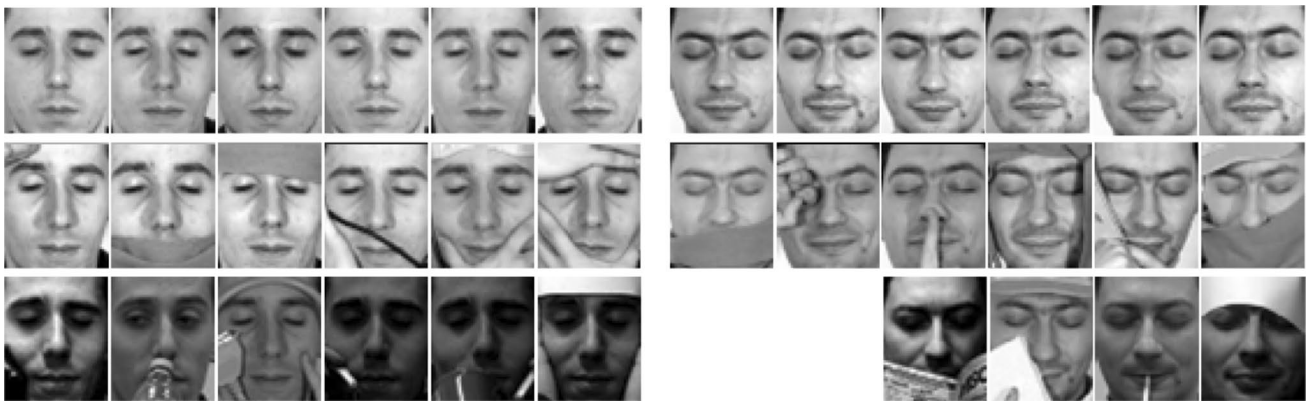


Fig. 8 The training and test samples of subject 1 (left) and subject 9 (right) from the UMB-DB database. Top row: the clean training images with normal illumination. Middle row: images from test set 1 containing various disguises and normal illuminations. Bottom row: images from test set 2 containing various disguises and bad illuminations

(about 8 for each subject), as shown in Fig. 5a. For testing, we construct 3 test sets. Test Set 1 selects $119 \times 3 = 357$ face images with only illumination variations (left light on, right light on, and all side lights on) from Session 1, as shown in Fig. 5b. Test Set 2 (or 3) selects 357 face images with scarf (or sunglasses) disguises mixed with three illumination variations (normal, left light on, and right light on) from Session 1, as shown in Fig. 5c, d. To test the performance of all compared methods against various feature dimensions, we downsample all training and test images into 4 pixel dimensions $14 \times 11 = 154$, $28 \times 23 = 644$, and $56 \times 46 = 2,576$, and $112 \times 92 = 10,304$, which correspond to downsampling ratios of 1/8, 1/4, 1/2, and 1, respectively.

Note that, from Test Set 1 to 3, the recognition difficulty is increasing. This is because the *aliasing* problem caused by occlusion becomes heavier from Test Set 1 to 3. Here, aliasing means there exists common structure between occlusion and training images. For example, the highlight illumination regions of face images in Fig. 5b might cause aliasing for the face images with pale color (as shown in Fig. 5e). In the AR database, the aliasing problems caused by highlight illuminations, scarves and sunglasses are successively increasing, as shown in Fig. 5e–g. Thus, the corresponding recognition difficulty from Test Set 1 to 3 is also increasing.

For the three test sets of the AR database, Fig. 6 compares the recognition performance of the 5 IGO-SECs versus different iterations (top row) and various feature dimensions (bottom row). According to the top row, with the hardness of recognition increasing, we have two observations: (1) the convergences of IGO-SECs, except for IGO-SEC-1, become worse (especially on Test Set 3), and (2) the performance gaps between IGO-SEC-1 and IGO-SEC-2 are shrinking, whereas the performance gaps between IGO-SEC-3 and IGO-SEC-4 become large. The former observation shows that all of the 3 factors (w, $\hat{y}$ and CC) are very important for the convergence of IGO-SEC, and the latter one verifies that the structural error metric CC contributes more than CEM. As per the bottom row of Fig. 6, with the recognition difficulty increasing, IGO-SEC-4 and IGO-SEC-5 come to have significant performance degeneration especially at the feature dimension 2,576 and on Test Set 3. This means that w and $\hat{y}$ are two important items to avoid performance degeneration.

By comparing Figs. 6 and 3, we can conclude that the performances of the IGO-SECs are roughly decreasing from IGO-SEC-1 to IGO-SEC-5. The only exception is on Test Set 1 (only with illumination variations) of AR, where IGO-SEC-3 outperforms IGO-SEC-2 at the feature dimensions 154 and 2,576. This means weight tuning might be more important than using an advanced structural error metric if there is no explicit occlusion.

Figure 7 and Table 3 compare the recognition performance of IGO-SEC with other 8 compared methods on the three test sets of AR. As can be seen, IGO-SEC-1 achieves the optimal recognition rates almost under all test cases. The only exception is at the highest dimension 10,304 on Test Set 3, where IGO-SEC-1 has a slight performance degradation and PCANet outperforms IGO-SEC-1 by 6.4%. Actually, for the three test sets of AR, most of the compared methods (especially for the five robust classifiers) have such performance degeneration. For example, on Test Set 1, CRC achieves its maximum recognition rate, 79.99%, at the pixel dimension 644 but it drops by 21.29% at 10,304. Clearly, such a performance degeneration is caused by the *curse of dimension*: with the feature dimension increasing, the challenging disguise also gains a large size and thus forms very significant local features which are more likely to incur aliasing. To deal with the problem incurred by the curse of dimension, a natural way is to detect and

Table 4 Recognition rates (%) on three test sets with various feature dimensions (FD) of the UMB-DB database

	Test set 1				Test set 2			
FD	154	644	2576	10304	154	644	2576	10304
CRC+IGO	56.70	69.55	81.28	84.64	53.80	63.59	75.54	81.52
CESR+IGO	70.67	78.77	88.27	91.90	60.33	68.48	82.07	85.87
RSC+IGO	70.67	79.61	88.55	92.18	61.41	70.11	82.07	85.87
NMR+IGO	60.06	88.83	90.22	91.06	43.48	66.85	84.78	86.96
SSEC+IGO	61.47	82.44	89.68	90.15	55.98	66.30	78.80	85.33
NN+PCANet	69.83	81.56	89.11	**96.09**	58.15	68.48	82.07	89.67
CRC+VGG	34.92	52.51	72.07	79.89	10.33	17.93	46.20	59.78
CRC+LCNN	74.86	81.28	90.78	94.13	37.50	54.89	79.35	87.50
IGO-SEC-1	**83.52**	**88.55**	**92.46**	92.46	**70.11**	**77.17**	**85.33**	**90.76**

The best result in each column is in bold

eliminate occlusion, such as SSEC and IGO-SEC. As seen from Table 3, compared to the other four robust classifiers (CRC, CESR, RSC and NMR) without using occlusion detection, SSEC at least postpones its peak recognition rate to the feature dimension 2,576 on Test Set 1 and has very weak performance degradation (also greatly outperforms the other methods) on Test Set 2. On the other hand, the 3 deep learning methods (PCANet, VGG and LCNN) almost do not have performance degeneration. But the reasons are various. PCANet actually solve the curse of dimension by using a much higher feature dimension (about 40 times of the pixel dimension). Such a high-dimensional feature is large enough to attenuate the negative influence incurred by occlusion. Things are totally different for VGG and LCNN. VGG and LCNN do not have a performance degeneration because they require using fixed input dimensions. In sum, to avoid performance degradation, there might exist various methods, but IGO-SEC might provide an optimal one, as demonstrated in Fig. 7.

8.4 Real-world occlusions using the UMB-DB database

We now evaluate the robustness of our algorithm against various real disguises using the UMB-DB face database. The evaluation protocols are identical to (Li et al. 2017). We transform all facial images into gray scales and crop and resize them into 112×92 pixel resolutions. For training, we use 433 non-occluded images with normal illuminations of the143 subjects. For testing, we consider two separate test sets of the 143 subjects. Test Set 1 contains 358 images of the whole subjects with various occlusions but normal illuminations, while Test Set 2 contains 184 images of the whole subjects with various occlusions but bad lighting conditions. Figure 8 gives some training and test examples of Subject 1 and Subject 9 from the UMB-DB database.

To test the recognition performances of the compared methods under various feature dimensions, we also downsample the training and test images into 154, 644, 2,576 and 10,304, which correspond to downsampling ratios of 1/8, 1/4, 1/2 and 1, respectively. The recognition rates are shown in Table 4. Different with the recognition results shown in Fig. 3 or Table 7, the recognition rates listed in Table 4 show that there is no performance degeneration on the two test sets of UMB-DB, although UMB-DB contains more complex occlusions than AR. This is mainly because the recognition tasks on the two test sets of UMB-DB are much easier than those on AR, as can be seen by comparing the data listed in Tables 4 and 7. The performance gaps between the two datasets further further show that the main challenge caused by occlusion dose not lie in its complexity, but lies in the aliasing problems (see Sect. 8.3) caused by occlusion. Indeed, in UMB-DB, it is not easy to find similar structures simultaneously implied in the occlusions and in the non-occluded training images.

On the other hand, Table 4 also indicates that IGO-SEC-1 achieves the optimal recognition performances almost for all test cases. The only exception is at the feature dimension 10,304 on Test Set 1, where PCANet, once again, obtains the optimal recognition rate by outperforming IGO-SEC-1 with 3.63%. As indicated in Sect. 8.3, the main reason lies in the very high-dimensional feature produced by PCANet. In addition, it is also worth noticing the good performances of NMR on the two test sets. Without using any occlusion detection and noise elimination, NMR achieves comparable results with IGO-SEC at the two highest feature dimensions (2,576 and 10,304). However, if we use the pixel feature instead of IGO, the recognition performances of NMR drop sharply. This means that what is critical for face recognition with occlusion is to accurately describe the distribution of the error caused by noises. To this end, seeking for a suitable robust feature space becomes very important but was not explored by the series works of Yang *et al.* (Luo et al.

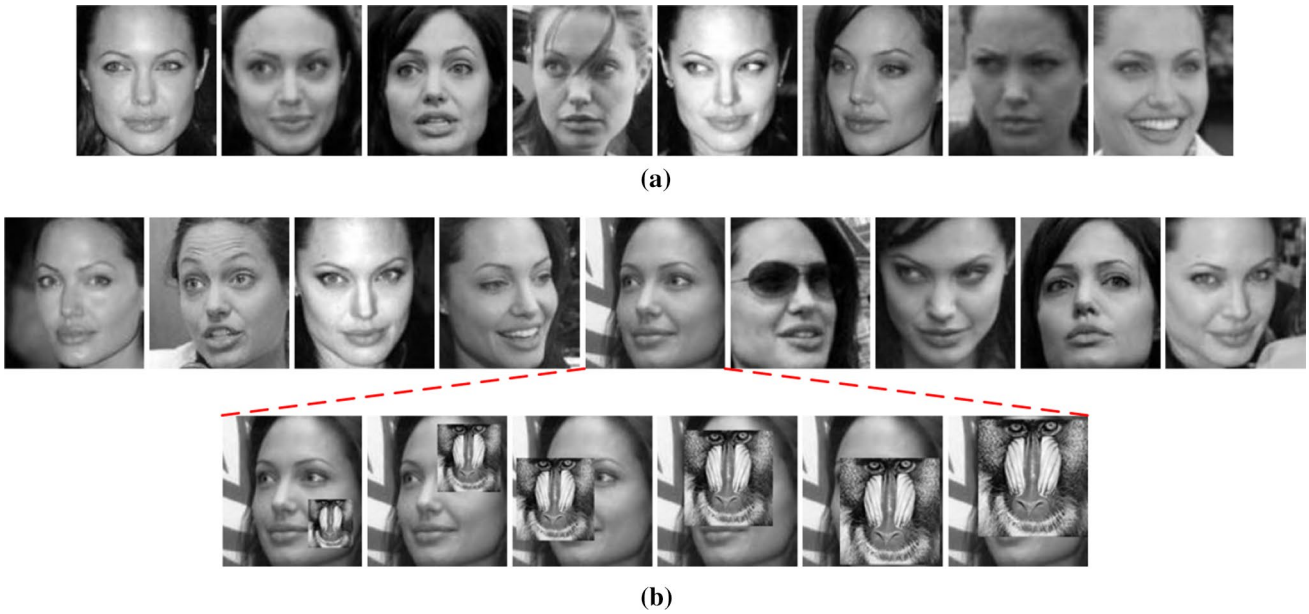


Fig. 9 The training and test samples from the LFW database (Huang et al. 2007). **a** Training set: each subject contains eight training samples with normal illumination conditions and without occlusion. **b** test set: each test image has a big pose variation and contains 0–60% synthetic occlusion (some test images also include natural occlusion)

Table 5 Recognition rates (%) with various occlusion percents (OP) on the test set of the LFW Database

OP	0%	10%	20%	30%	40%	50%	60%
CRC+IGO	60.57	52.41	43.08	33.24	19.86	11.22	4.88
CESR+IGO	50.15	46.90	40.31	32.73	23.14	15.82	8.93
RSC+IGO	58.67	52.04	42.16	30.69	20.01	10.31	5.32
NMR+IGO	63.85	57.40	48.29	38.37	24.78	14.21	6.38
SSEC+IGO	54.05	49.45	29.66	24.05	18.33	12.50	8.86
NN+PCANet	67.41	56.71	50.77	42.67	34.18	24.64	16.87
CRC+VGG	**98.07**	88.48	62.72	29.05	8.24	2.37	1.31
CRC+LCNN	96.10	**88.74**	**68.99**	43.48	18.80	5.98	1.86
IGO-SEC-1	73.37	68.47	60.71	**52.73**	**43.37**	**32.76**	**21.32**

The best result in each column is in bold

2015; Qian et al. 2015; Yang et al. 2017). This problem will be left as our future work.

8.5 Unconstrained face recognition with synthetic occlusion using the LFW database

In this section, we evaluate the effectiveness of our algorithm against occlusion in unconstrained settings using the LFW database. The evaluation protocol is identical to (Li et al. 2017). Specifically, we use a subset of LFW, which consists of 217 subjects with no less than 8 samples per subject, from the aligned LFW (LFW-deepfunneled) (Huang et al. 2012). In this subset, we select 8 samples with fewer pose and illumination variations from each subject for training and the left images are used for testing. In order to test the robustness of the compared methods against occlusion, we also replace a random block of each test image with a "baboon" image which occupies 0% to 60% (with step size 10%) of the test image. Some training and test examples are shown in Fig. 9. All images are cropped and resized to 62×58 pixels.

Table 5 reports the recognition rates of IGO-SEC with other 8 compared methods on the test set of the LFW dataset. As can be seen, an obvious result we can first achieve is that when the occlusion level is very small ($\leq 20\%$), the classical deep learning methods (VGG and LCNN) dramatically outperform the other compared approaches. Specifically, CRC+VGG and CRC+LCNN outperform the other methods (besides IGO-SEC-1) by 30% or so and surpass our proposed IGO-SEC-1 by about 20%. However, once the occlusion percent is larger

than 20%, the recognition rates of the two CNN methods drop sharply, while the proposed IGO-SEC-1 keep more stable recognition performances and outperform the other methods. This clearly demonstrates the powerful occlusion robustness of our proposed structural error coding method.

9 Conclusions

Real-world face recognition systems are challenged by many mixed variations. To effectively perform face recognition with occlusion and other coexisting common variations like illumination changes and Gaussian white noises, we suggest to embed the Image Gradient Orientations (IGO) into robust error coding models. By analyzing the distribution of the reconstruction error between the occluded face image and its reconstruction in the occluded and non-occluded regions respectively, we propose a joint probabilistic generative model for a novel IGO-embedded Structural Error Coding (IGO-SEC) model. Around how to fully use the potential robustness of the IGO features and how to enhance the quality of the error image in IGO-SEC, we further present a new reconstruction method by integrating the advantages of the pixel and IGO features, and a new robust structural error metric by fusing the spatial structure and statistical information of the error image in IGO. Extensive experiments on 4 benchmark face databases widely verify the robustness of the proposed IGO-SEC model on various occlusion levels and feature dimensions. How to borrow the idea of robust error coding models into the deep learning schemes will be left as our future work.

Acknowledgements This work is partially supported by Natural Science Foundation of Zhejiang Province (LY18F020031), National Natural Science Foundation of China (61379020, 61402411, 61802347).

References

Ahonen T, Hadid A, Pietikainen M (2006) Face description with local binary patterns: application to face recognition. IEEE Trans Pattern Anal Mach Intell 28:2037–2041

Chan TH, Jia K, Gao S, Lu J, Zeng Z, Ma Y (2015) PCANet: a simple deep learning baseline for image classification? IEEE Trans Image Process 24:5017–5032

Chen SS, Donoho DL, Saunders MA (2001) Atomic decomposition by basis pursuit. SIAM Rev 43:129–159

Colombo A, Cusano C, Schettini R (2011) UMB-DB: a database of partially occluded 3d faces. In: IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, pp 2113–2119

Georghiades AS, Belhumeur PN, Kriegman DJ (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23:643–660

Ghazi MM, Ekenel HK (2016) A comprehensive analysis of deep learning based representation for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA, pp 34–41

He R, Zheng W, Hu B (2011) Maximum correntropy criterion for robust face recognition. IEEE Trans Pattern Anal Mach Intell 33:1561–1576

He R, Zheng WS, Tan T, Sun Z (2014) Half-quadratic-based iterative minimization for robust sparse representation. IEEE Trans Pattern Anal Mach Intell 36:261–275

Hua G, Yang MH, Learned-Miller E, Ma Y, Turk M, Kriegman DJ, Huang TS (2011) Introduction to the special section on real-world face recognition. IEEE Trans Pattern Anal Mach Intell 33:1921–1924

Huang GB, Mattar M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07–49, University of Massachusetts

Huang GB, Mattar M, Lee H, Learned-Miller E (2012) Learning to align from scratch. Advances in neural information processing systems. Lake Tahoe, Nevada, USA, pp 764–772

Kolmogorov V, Zabih R (2004) What energy functions can be minimized via graph cuts? IEEE Trans Pattern Anal Mach Intell 26:147–159

Lee K, Ho J, Kriegman D (2005) Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans Pattern Anal Mach Intell 27:684–698

Li XX, Dai DQ, Zhang XF, Ren CX (2013) Structured sparse error coding for face recognition with occlusion. IEEE Trans Image Process 22:1889–1900

Li XX, He L, Hao P, Liu Z, Li J (2017) Adaptive weberfaces for occlusion-robust face representation and recognition. IET Image Process 11:964–975

Liang R, Li XX (2015) Mixed error coding for face recognition with mixed occlusions. In: International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, pp 3657–3663

Lin D, Tang X (2007) Quality-driven face occlusion detection and recovery. In: IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, pp 1–7

Liu W, Pokharel PP, Principe JC (2007) Correntropy: properties and applications in non-gaussian signal processing. IEEE Trans Signal Process 55:5286–5298

Luo L, Yang J, Qian J, Tai Y (2015) Nuclear-l1 norm joint regression for face reconstruction and recognition with mixed noise. Pattern Recog 48:3811–3824

Martinez AM (1998) The AR face database. Technical Report 24, Computer Visual Center, Ohio State University

Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: British Machine Vision Conference, London: BMVA Press, Swansea, UK, pp 41.1–41.12

Portugal L, Judice J, Vicente L (1994) A comparison of block pivoting and interior-point algorithms for linear least squares problems with nonnegative variables. Math Comput 63:625–644

Qian J, Yang J, Xu Y (2013) Local structure-based image decomposition for feature extraction with applications to face recognition. IEEE Trans Image Process 22:3591–3603

Qian J, Luo L, Yang J, Zhang F, Lin Z (2015) Robust nuclear norm regularized regression for face recognition with occlusion. Pattern Recog 48:3145–3159

Sun Y, Wang X, Tang X (2015) Deeply learned face representations are sparse, selective, and robust. In: IEEE Conference on Computer Vision and Pattern Recognition. MA, USA, Boston, pp 2892–2900

Sun Y, Wang X, Tang X (2016) Sparsifying neural network connections for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, pp 4856–4864

Taigman Y, Yang M, Ranzato M, Wolf L (2015) Web-scale training for face identification. In: IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, pp 2746–2754

Tzimiropoulos G, Zafeiriou S, Pantic M (2012) Subspace learning from image gradient orientations. IEEE Trans Pattern Anal Mach Intell 34:2454–66

Wang JJY, Wang (2013) Non-negative matrix factorization by maximizing correntropy for cancer clustering. BMC Bioinform 14:1–11

Wei X, Li CT, Hu Y (2012) Robust face recognition under varying illumination and occlusion considering structured sparsity. In: International conference on digital image computing techniques and applications. Fremantle, Australia, pp 1–7

Wright J, Ma Y (2010) Dense error correction via ℓ^1-minimization. IEEE Trans Inf Theory 56:3540–3560

Wright J, Yang A, Ganesh A, Sastry S, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31:210–227

Wu X, He R, Sun Z (2015) A lightened CNN for deep face representation. CoRR abs/1511.02683. arXiv:1511.02683

Yang J, Luo L, Qian J, Tai Y, Zhang F, Xu Y (2017) Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. IEEE Trans Pattern Anal Mach Intell 39:156–171

Yang M, Zhang L, Yang J, Zhang D (2011) Robust sparse coding for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, pp 625–632

Yang M, Zhang L, Shiu SC, Zhang D (2013a) Gabor feature based robust representation and classification for face recognition with gabor occlusion dictionary. Pattern Recog. 46:1865–1878

Yang M, Zhang L, Shiu SK, Zhang D (2013b) Robust kernel representation with statistical local features for face recognition. IEEE Trans Neural Netw Learn Syst 24:900–912

Yang M, Zhang L, Yang J, Zhang D (2013c) Regularized robust coding for face recognition. IEEE Trans Image Process 22:1753–1766

Zhang L, Yang M, Feng X (2011) Sparse representation or collaborative representation: Which helps face recognition? In: IEEE International Conference on Computer Vision, Barcelona, Spain, pp 471–478

Zhang T, Tang YY, Fang B, Shang Z, Liu X (2009) Face recognition under varying illumination using gradientfaces. IEEE Trans Image Process 18:2599–2606

Zhao F, Feng J, Zhao J, Yang W, Yan S (2018) Robust lstm-autoencoders for face de-occlusion in the wild. IEEE Trans Image Process 27:778–790

Zhou Z, Wagner A, Mobahi H, Wright J, Ma Y (2009) Face recognition with contiguous occlusion using markov random fields. In: IEEE International Conference on Computer Vision. Kyoto, Japan, pp 1050–1057