Dynamic Image-to-Class Warping for Occluded Face Recognition

Xingjie Wei, Chang-Tsun Li, *Senior Member, IEEE*, Zhen Lei, *Member, IEEE*, Dong Yi, and Stan Z. Li, *Fellow, IEEE*

***Abstract*— Face recognition (FR) systems in real-world applications need to deal with a wide range of interferences, such as occlusions and disguises in face images. Compared with other forms of interferences such as nonuniform illumination and pose changes, face with occlusions has not attracted enough attention yet. A novel approach, coined dynamic image-to-class warping (DICW), is proposed in this work to deal with this challenge in FR. The face consists of the forehead, eyes, nose, mouth, and chin in a natural order and this order does not change despite occlusions. Thus, a face image is partitioned into patches, which are then concatenated in the raster scan order to form an ordered sequence. Considering this order information, DICW computes the image-to-class distance between a query face and those of an enrolled subject by finding the optimal alignment between the query sequence and all sequences of that subject along both the time dimension and within-class dimension. Unlike most existing methods, our method is able to deal with occlusions which exist in both gallery and probe images. Extensive experiments on public face databases with various types of occlusions have confirmed the effectiveness of the proposed method.**

***Index Terms*— Face recognition, occlusion, image-to-class distance, dynamic time warping, biometrics.**

I. Introduction

FACE recognition (FR) is one of the most active research topics in computer vision and patten recognition over the past few decades. Nowadays, automatic FR system achieves significant progress in controlled conditions. However, the performance in unconstrained conditions (e.g., large variations in illumination, pose, expression, etc.) is still unsatisfactory. In the real-world environments, faces are easily occluded by facial accessories (e.g., sunglasses, scarf, hat, veil), objects in front of the face (e.g., hand, food, mobile phone), extreme illumination (e.g., shadow), self-occlusion (e.g., non-frontal pose) or poor image quality (e.g., blurring). The difficulty of occluded FR is twofold. Firstly, occlusions distort the discriminative facial features and increases the distance between two face images of the same subject in the feature space. The intra-class variations are larger than the inter-class variations, which results in poorer recognition performance. Secondly, when facial landmarks are occluded, large registration errors usually occur and degrade the recognition rate [1].

Note that there are two related but different problems to FR with occlusions: *occluded face detection* and *occluded face recovery*. The first task is to determine whether a face image is occluded or not [2], which can be used for automatically rejecting the occluded images in applications such as passport image enrolment. This rejection mechanism is not always applicable in some scenarios (e.g., surveillance) where no alternative image can be obtained due to the lack of user cooperation. The second task is to restore the occluded region in face images [3], [4]. It can recover the occluded area but is unable to directly contribute to recognition since the identity information can be contaminated during inpainting.

An intuitive idea for handling occlusions in FR is to detect the occluded region first and then perform recognition using only the unoccluded part. Min *et al.* [5] adopted a SVM classifier to detect the occluded region in a face image then used only the unoccluded area of a probe face (i.e., query face) as well as the corresponding area of the gallery faces (i.e., reference faces) for recognition. But note that the occlusion types in the training images are the same as those in the testing images. Jia and Martinez [6], [7] used a skin colour based mask to remove the occluded area for recognition. However, the types of occlusions are unpredictable in practical scenarios. The location, size and shape of occlusions are unknown, hence increasing the difficulty in segmenting the occluded region from the face images. Currently most of the occlusion detectors are trained on faces with specific types of occlusions (i.e., the training is data-dependent) and hence generalise poorly to various types of occlusions in the real-world environment.

In this paper, we focus on performing recognition in the presence of occlusions. There are two main categories of approaches in this direction. The first is the *reconstruction based approaches* which treat occluded FR as a reconstruction problem [6], [8]–[13]. The sparse representation based classification (SRC) [8] is a representative example. A clean image is reconstructed from an occluded probe image by a linear combination of gallery images. Then the occluded image is assigned to the class with the minimal reconstruction error.

Manuscript received March 15, 2014; revised July 11, 2014; accepted September 3, 2014. Date of publication September 22, 2014; date of current version November 10, 2014. The work of Z. Lei was supported by the National Natural Science Foundation of China under Project 61103156 and Project 61473291. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gérard Medioni. *(Corresponding author: Chang-Tsun Li.)*

X. Wei and C.-T. Li are with the Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: x.wei@warwick.ac.uk; ct.li@warwick.ac.uk).

Z. Lei, D. Yi, and S. Z. Li are with the Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zlei@cbsr.ia.ac.cn; dyi@cbsr.ia.ac.cn; szli@cbsr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2014.2359632

TABLE I
THREE TYPICAL OCCLUSION CASES

	Gallery	Probe	Scenario	Application
Uvs.O:	Unoccluded	Occluded	Access control, ID management	Facility access, immigration, user registration, online authentication
Ovs.U:	Occluded	Unoccluded	Law enforcement, security, surveillance	Suspect investigation, shoplifter recognition, criminal face retrieval, terrorist alert, CCTV control
Ovs.O:	Occluded	Occluded		

The reconstruction based approaches usually require a large number of samples per subject to represent a probe image. However, a sufficient number of samples are not always available in practical scenarios.

The second category is the *local matching based approaches* [14]–[17]. Facial features are extracted from local areas of a face, for example, overlapping or non-overlapping patches of an image, so the affected and unaffected parts of the face can be analysed in isolation. In order to minimise matching errors due to occluded parts, different strategies such as local subspace learning [14], [15], partial distance learning [16] and multi-task sparse representation learning [17] are performed. Our method belongs to this category but does not require training.

In addition to the above approaches, which focus on improving the robustness during the recognition stage, recently many researchers [18], [19] also pay attention to the image presentation stage and attempt to extract stable, occlusion-insensitive features from face images. Since the forms of occlusions in real-world scenarios are unpredictable, it is still difficult to find a suitable representation which is insensitive to the variations in occlusions.

Most of the current methods assume that occlusions only exist in the probe images and the gallery or training images are *clean*. In practical scenarios, occlusions may occur in both gallery and probe images [6], [7], [20]. When the number of gallery/training images is limited, excluding these occluded images would, on the one hand, lead to *small sample size (SSS)* problem [21], and on the other hand, ignore useful information for recognition [20]. We summarise three occlusion cases in Table I, which a FR system may encounter in the real-world applications. Most of the current methods rely on a clean gallery or training set and only consider the first case. The latter two cases would also occur in real environment but have not yet received much attention.

We propose a local matching based method, Dynamic Image-to-Class Warping (DICW), for occluded FR. DICW is motivated by the Dynamic Time Warping (DTW) algorithm [22] which allows elastic match of two time sequences. It has been successfully applied to the area of speech recognition [22]. In our work, an image is partitioned into patches, which are then concatenated in the raster scan order to form a sequence. In this way, a face is represented by a patch sequence which contains the *order information* of facial features. DICW calculates the *Image-to-Class* distance between a query face and those of an enrolled subject by finding the optimal alignment between the query sequence and all enrolled sequences of that subject. Our method allows elastic match in both *time* and *with-class* directions.

Most of the existing works that simply treat occluded FR as a signal recovery problem or just employ the framework for general object classification, neglect the inherent structure of the face. Wang *et al.* proposed a Markov Random Field (MRF) based method [23] for FR and confirmed that the contextual information between facial features plays an important role in recognition. In this paper, we propose a novel approach that takes the *facial order*, which contains the geometry information of the face, into account when recognising partially occluded faces. In uncontrolled environments with uncooperative subjects, the occlusion preprocessing and the collection of sufficient and representative training samples are generally very difficult. Our method which performs recognition directly in the presence of occlusions and does not require training, is hence feasible for realistic FR applications.

This paper is built upon our preliminary work reported in [24] and [25]. The remainder of this paper is organised as follows. The proposed Dynamic Image-to-Class Warping method, from image representation, modelling to implementation, is described in Section II. Extensive experiments including discussions are presented in Section III. Further analysis about why the proposed method works; when and why it will fail and how to improve it is discussed in Section IV. Finally the work is concluded in Section V.

II. DYNAMIC IMAGE-TO-CLASS WARPING

A. Image Representation

An image is partitioned into J non-overlapping patches of $d \times d'$ pixels. Those patches are then concatenated in the raster scan order (i.e., from left to right and top to bottom) to form a single sequence. The reason for doing so is that the forehead, eyes, nose, mouth and chin are located in the face in a natural order, which does not change despite occlusions or imprecise registration. This *spatial facial order*, which is contained in the patch sequence, can be viewed as the *temporal order* in the time sequence. In this way, a face image can be viewed as a time sequence so the image matching problem can be handled by the time series analysis technique like DTW [22].

Let $f(x, y)$ be the intensity of the pixel at coordinates (x, y) and $f_j(x, y)$ be the j-patch. A difference patch $\triangle f_j(x, y)$ is computed (Fig. 1) by subtracting $f_j(x, y)$ from its immediate

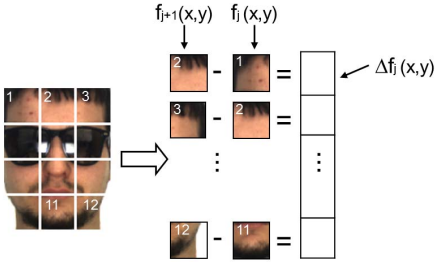


Fig. 1. Image representation of DICW.

neighbouring patch $f_{j+1}(x, y)$ as:

$$\triangle f_j(x, y) = f_{j+1}(x, y) - f_j(x, y) \tag{1}$$

where $j \in \{1, 2, \ldots, J-1\}$. Note that here the length of the *difference patch* sequence is $J-1$.

A difference patch $\triangle f_j(\cdot)$ actually can be viewed as the approximation of the first-order derivative of adjacent patch $f_{j+1}(\cdot)$ and $f_j(\cdot)$. The salient facial features which represent textured regions such as eyes, nose and mouth can be enhanced since the first-order derivative operator is sensitive to edges.

We use 3,200 occluded-unoccluded image pairs of the same class and different classes from the AR database [26], respectively (6,400 pairs in total) to calculate the image distance distributions.[1] As shown in Fig. 2, the distance distributions of the same and different classes are separated more widely when using the difference patches (Fig. 2b).

B. Modelling

Face matching is implemented by defining a distance measurement between sequences and using the distance as the basis for classification. Generally, a small distance is expected if two sequences are similar to each other. DICW is based on the classical DTW algorithm [22] which is used to compute the distance between two time sequences. Here we use an example to quickly illustrate the main idea of DTW (more details of the algorithm can be found in [22]). As shown in Fig. 3, there are two sequences (each digit indicates a data point):

$$A = (a_1, a_2, a_3, a_4, a_5) = (3, 1, 10, 3, 2)$$
$$B = (b_1, b_2, b_3, b_4, b_5) = (3, 1, 2, 10, 3).$$

The Euclidean distance (i.e., using point-wise matching, Fig. 3a) between them is $\sqrt{(a_1-b_1)^2 + \cdots + (a_5-b_5)^2} = \sqrt{0+0+64+49+1} \approx 10.68$ which is a bit large for these two similar sequences. However, if we *warp* these two sequences in a non-linear way by shrinking or expanding them along the time axis during matching (i.e., allows flexible correspondences), the distance between A and B can be largely reduced[2] to 2 (Fig. 3b). DTW, which is based on this idea, calculates the distance between two time sequences by finding the optimal alignment between them with the minimal overall cost. This will help to reduce the distance error caused by some *noise* data points and ensure that the distance between similar sequences is relatively small. In addition, the *temporal order* is considered during matching, thus cross-matching (which reverses the order of data points) is not allowed even it can lead to shorter distance (Fig. 3c). Especially for FR, this is reasonable since the order of facial features should not be turned back.

Adopting this idea for FR, we want to find the optimal alignment between face sequences while minimising the distance caused by occluded patches. In this work, instead of finding the alignment between two sequences, we seek the alignment between a sequence and the sequence *set* of a given class (i.e., subject). A probe image consisting of M patch features is denoted by $\boldsymbol{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m, \ldots, \boldsymbol{p}_M)$. Here $\boldsymbol{P}$ is an ordered list where each element $\boldsymbol{p}_m$ is a patch feature vector (e.g., $\triangle f(\cdot)$ in Section II-A). The gallery set of a given class containing K images is denoted by $\boldsymbol{G} = \{\boldsymbol{G}_1, \ldots, \boldsymbol{G}_k, \ldots, \boldsymbol{G}_K\}$. The k-th gallery image is similarly represented as a sequence of N patch features as $\boldsymbol{G}_k = (\boldsymbol{g}_{1k}, \ldots, \boldsymbol{g}_{nk}, \ldots, \boldsymbol{g}_{Nk})$ where $\boldsymbol{g}_{nk}$ represents a patch feature vector like $\boldsymbol{p}_m$. Note that the number of patches in the probe image and that in the gallery image can be different (i.e., the values of M and N can be different) since the DTW model is able to deal with sequences with different lengths [22].

A warping path $\boldsymbol{W}$ indicating the matching correspondence of patches between $\boldsymbol{P}$ and $\boldsymbol{G}$ with T warping steps in time axis is defined as $\boldsymbol{W} = (w(1), \ldots, w(t), \ldots, w(T))$ with:

$$w(t) = (m, n, k) : \{1, 2, \ldots, T\} \rightarrow \{1, 2, \ldots, M\} \times \{1, 2, \ldots, N\} \times \{1, 2, \ldots, K\} \tag{2}$$

where $\times$ indicates the Cartesian product operator and $\max\{M, N\} \leqslant T \leqslant M + N - 1$. $w(t) = (m, n, k)$ is a triple indicating that patch $\boldsymbol{p}_m$ is matched to patch $\boldsymbol{g}_{nk}$ at step t.

Similar to the DTW model [22], $\boldsymbol{W}$ in DICW satisfies the following four constraints:

1) Boundary: $w(1) = (1, 1, k)$ and $w(T) = (M, N, k')$. The path starts at matching $\boldsymbol{p}_1$ to $\boldsymbol{g}_{1k}$ and ends at matching $\boldsymbol{p}_M$ to $\boldsymbol{g}_{Nk'}$. Note that no restrictions are placed on k and k'. From step 1 to T, k and k' can be any value from 1 to K since the probe patch can be matched with patches from all K gallery images.
2) Monotonicity: Given $w(t) = (m, n, k)$, the preceding triple $w(t-1) = (m', n', k')$ satisfies that $m' \leqslant m$ and $n' \leqslant n$. The warping path preserves the *temporal order* and increase monotonically.
3) Continuity: Given $w(t) = (m, n, k)$, the preceding triple $w(t-1) = (m', n', k')$ satisfies that $m - m' \leqslant 1$ and $n - n' \leqslant 1$. The indexes of the path increase by 1 in each step, which means that each step makes smooth transitions along the *time* dimension.
4) Window constraint: Given $w(t) = (m, n, k)$, it satisfies $|m - n| \leqslant l$ where $l \in \mathbb{N}^+$ is the window width [22]. The window constraint is designed to reduce the computational cost of DICW. But it is also meaningful for the specific FR problem since a probe patch (e.g., eye) should not match to a patch (e.g., mouth) too far away. The window with a width l is able to constrain the warping path within an appropriate range.

[1] We use Euclidean distance as measurement. The image size is 83×60 pixels and the patch size is 5×5 pixels.

[2] Computation details see [22].

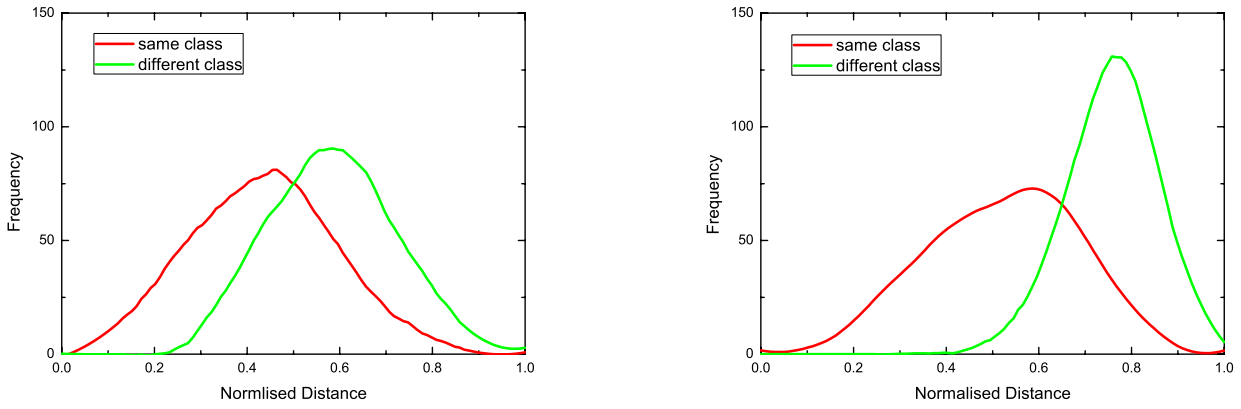


Fig. 2. Distributions of face image distance of the same and different classes. Using the difference patch (b), the distance distribution of the same class and that of the different classes are separated more widely compared with those using the original patch (a).

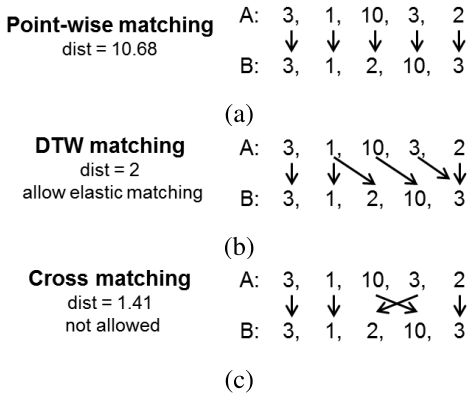


Fig. 3. Various ways of sequence matching. a) Point-wise matching, b) DTW matching, and c) cross matching.

These constraints are extended from the constraints of the DTW algorithm. However, they are also very meaningful in the context of FR with the image representation defined in Section II-A. Our method represents a face image as a patch sequence thus here the image matching problem can be solved by the time series analysis technique.

In order to explain the concept of *warping path*, we take the aforementioned sequences A and B as an example. In Fig. 4a, each grid on the right hand side indicates a possible matching correspondence. The indexes of the red grids indicate the matching between A and B by DTW (i.e., the optimal warping path with the minimal matching cost) as shown in the left part (here $T = 6$). Likewise, the same procedure of DICW is shown in Fig. 4b. Compared with DTW, an additional index is added in the warping step of DICW to index different gallery sequences. In this way, the *warping* is performed in two directions: 1) a probe sequence $\boldsymbol{P}$ is aligned to a set of gallery sequences $\boldsymbol{G}$ according to the *time* dimension (maintaining the *facial order*) and 2) simultaneously, at each warping step, each patch in $\boldsymbol{P}$ can be matched with any patch among all gallery sequences along the *within-class* dimension. Our method allows elastic match in both of the aforementioned two directions.

We define the local distance [22] $C_{m,n,k} = d(\boldsymbol{p}_m, \boldsymbol{g}_{nk})$ as the distance between two patches $\boldsymbol{p}_m$ and $\boldsymbol{g}_{nk}$. $d(\cdot)$ can be any distance measurement such as the Euclidean distance or the Cosine distance. The overall matching cost of $\boldsymbol{W}$ is the sum of the local distance of each warping step:

$$S(\boldsymbol{W}) = \sum_{t=1}^{T} C_{w_t} \tag{3}$$

The optimal warping path $\boldsymbol{W}^*$ (i.e., the red grid path in Fig. 4b) is the path that minimises $S(\boldsymbol{W})$. The *Image-to-Class* distance between $\boldsymbol{P}$ and $\boldsymbol{G}$ is simply the overall cost of $\boldsymbol{W}^*$:

$$dist_{DICW}(\boldsymbol{P}, \boldsymbol{G}) = \min_{\boldsymbol{W}} \sum_{t=1}^{T} C_{w_t} \tag{4}$$

After computing $dist_{DICW}$ between $\boldsymbol{P}$ and each enrolled subject in the database, a classifier such as the Nearest Neighbour classifier can be adopted for classification based on $dist_{DICW}$.

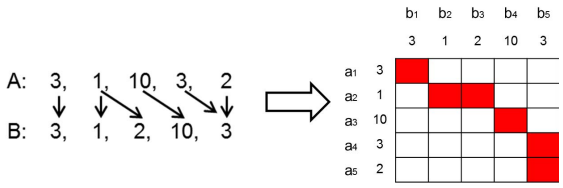
C. Implementation Through Dynamic Programming

To compute $dist_{DICW}(\boldsymbol{P}, \boldsymbol{G})$ in (4), one could test every possible warping path but with a high computational cost. Fortunately, (4) can be solved efficiently by *Dynamic Programming*. A three-dimensional matrix $\boldsymbol{D} \in \mathbb{R}^{M \times N \times K}$ is created to store the cumulative distance. The element $D_{m,n,k}$ stores the cost of the optimal warping path of matching the first m probe patches to the set of first n patches of each gallery sequence and at the same time the m-th patch $\boldsymbol{p}_m$ is matched to the patch from the k-th gallery image. The calculation of the final optimal cost $dist_{DICW}(\boldsymbol{P}, \boldsymbol{G})$ is based on the results of a series of predecessors. $\boldsymbol{D}$ can be computed recursively as:

$$D_{m,n,k} = \min \left\{ \begin{array}{l} \boldsymbol{D}_{\{(m-1,n-1)\}\times\{1,2,\ldots,K\}}, \\ \boldsymbol{D}_{\{(m-1,n)\}\times\{1,2,\ldots,K\}}, \\ \boldsymbol{D}_{\{(m,n-1)\}\times\{1,2,\ldots,K\}} \end{array} \right\} + C_{m,n,k} \tag{5}$$

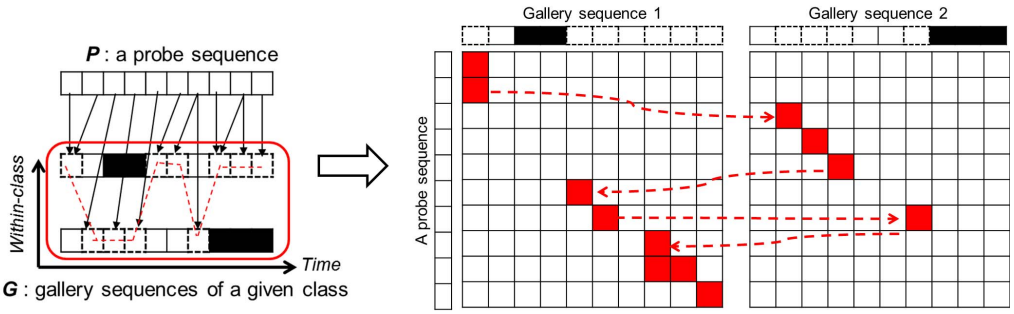
where the initialisation is done by extending $\boldsymbol{D}$ as an $(M+1) \times (N+1) \times K$ matrix and setting $\boldsymbol{D}_{0,0,\cdot} = 0$, $\boldsymbol{D}_{0,n,\cdot} = \boldsymbol{D}_{m,0,\cdot} = \infty$. Thus, $dist_{DICW}(\boldsymbol{P}, \boldsymbol{G})$ can be obtained as follows:

$$dist_{DICW}(\boldsymbol{P}, \boldsymbol{G}) = \min_{k \in \{1,2,\ldots,K\}} \{\boldsymbol{D}_{M,N,k}\} \tag{6}$$



$$\boldsymbol{W} = ((1,1),(2,2),(2,3),(3,4),(4,5),(5,5)), T = 6$$

(a)



$$\boldsymbol{W} = ((1,1,1),(2,1,1),(3,2,2),(4,3,2),(5,4,2),(6,5,1),(7,6,1),(7,7,2),(8,8,1),(9,8,1),(9,9,1),(10,10,1)), T = 12$$

(b)

Fig. 4. An illustration of warping path in (a) DTW and the (b) proposed DICW. The arrows indicate the matching correspondence. The dashed line marks the optimal warping path. Black blocks indicate the occluded patches.

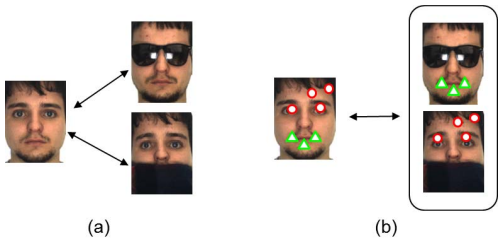


Fig. 5. The illustration of (a) the *Image-to-Image* and (b) the *Image-to-Class* comparison. Matched features are indicated by the same symbol.

Different from the point-wise matching (here each patch is viewed as a data point), our method tries every possible warping path under the temporal constraints then selects the one with minimal overall cost. So the warping path with large distance error will not be selected. The *Image-to-Class* distance is the *globally* optimal cost for matching. Although occlusions are not directly removed, avoiding large distance error by warping is helpful for classification from our experimental results (see Section III).

In addition, a patch of the probe image can be matched to patches of K different gallery images of the same class. Because the chance that all patches at the same location of the K images are occluded is low, the chance that a probe patch is compared to an unoccluded patch at the same location is thus higher. When occlusions occur in probe or/and gallery images, the *Image-to-Image* distance may be large. However, our model is able to exploit the information from different gallery images and reduce the effect of occlusions (Fig. 5). Algorithm 1 summarises the procedure of computing the *Image-to-Class* distance between a probe image and a class. l is the window width and usually set to 10% of $\max\{M, N\}$ [22]. Computational complexity is analysed in Section III-D6.

Algorithm 1 Dynamic Image-to-Class Warping Distance DICW($\boldsymbol{P}, \boldsymbol{G}, l$)

Input:
$\boldsymbol{P}$: a probe sequence with M patches;
$\boldsymbol{G}$: a set of K gallery sequences (each with N patches) of a given class;
l: the window width;

Output:
$dist_{DICW}$: the *Image-to-Class* distance between $\boldsymbol{P}$ and $\boldsymbol{G}$;

Set each element in $\boldsymbol{D}$ to ∞;
$\boldsymbol{D}[0, 0, 1:K] = 0$;
$l = \max\{l, |M - N|\}$;
Compute the local distance matrix $\boldsymbol{C}$;
for $m = 1$ to M **do**
  for $n = \max\{1, m-l\}$ to $\min\{N, m+l\}$ **do**
    minNeighbour=$\min \left\{ \begin{array}{l} \boldsymbol{D}[m-1, n-1, 1:K], \\ \boldsymbol{D}[m-1, n, 1:K], \\ \boldsymbol{D}[m, n-1, 1:K] \end{array} \right\}$;
    for $k = 1$ to K **do**
      $D[m, n, k]$ =minNeighbour+$C[m, n, k]$;
    end for
  end for
end for
$dist_{DICW} = \min\{\boldsymbol{D}[M, N, 1:K]\}$;
return $dist_{DICW}$;

III. Experimental Analysis

In this Section, we evaluate the proposed method using four databases (FRGC [27], AR [26], TFWM [28] and LFW [29]). We perform identification tasks according to the three cases (i.e., **Uvs.O**, **Ovs.U** and **Ovs.O**) described in Section I. We first consider the scenario with occlusions occur only in probe images (i.e., **Uvs.O**) and test our method using different number of gallery images per subject. We will demonstrate that our method works well even when a very limited number

of images are available for each subject. Next, we consider the situation that occlusions exist in gallery images, which is a case most of the current works do not take account. We fix the number of gallery images per subject and conduct experiments step by step: firstly the probe images are not occluded (i.e., **Ovs.U**); and then both the gallery and probe images are occluded (i.e., **Ovs.O**). Note that, for comparison purpose, the experiments on the FRGC and the AR databases also include the case that no occlusion is presented in both gallery and probe images to confirm that DICW is also effective in general conditions. In addition, we also extend DICW to verification tasks with faces containing large uncontrolled variations.

Note that in all experiments, the gallery image set is disjoint with all probe sets. Considering that the gallery and probe images are at the same scale, in the experiments, the probe images and the gallery images are partitioned into the same number of patches, i.e., $M = N$ as defined in Section II-B. As recommended in the work [30], the Euclidean distance and the Cosine distance are used as local distance measurements for the pixel intensity feature and the LBP feature [31], respectively.

We quantitatively compare DICW with some representative methods in the literature: the supervised linear SVM [32] using PCA [33] for feature extraction (PCA + LSVM), the reconstruction based SRC [8] as introduced in Section I, the *Image-to-Class* distance based Naive Bayes Nearest Neighbour (NBNN) [34] as ours, and the baseline, Hidden Markov models (HMM) [35] which also considers the order information in a face. We use the difference patch representation as defined in Section II-A in NBNN and DICW. For comparison purpose, we also report the results of using the original patche (referred to OP-NBNN and OP-Warp, respectively).

Note that NBNN is a local patch based method which also exploits the *Image-to-Class* distance. But it does not consider the spatial relationship between patches like ours. To improve the performance, a location weight α [34] is used in NBNN to constrain matching patches according to their locations. We tested different values of α and found that the performance of NBNN is highly dependent on the value of α and different testing data (e.g., different occlusion level, location) requires different value even within the same database. So we also reported the best result for each test with the optimal α value (as OP-NBNN-ub and NBNN-ub). The performance of OP-NBNN-ub and NBNN-ub can be seen as the upper bound of the performance of NBNN, which is a competitive comparison for DICW.

A. Face Identification With Randomly Located Occlusions

We first evaluate the proposed method using the Face Recognition Grand Challenge (FRGC) database [27] with randomly located occlusions. Note that in each image, the locations of occlusions are randomly chosen and unknown to the algorithm. Especially, in the **Ovs.O** scenario, the locations of occlusions in the gallery images are different from those in the probe images. We use these images with randomly located occlusions to evaluate the effectiveness of DICW when there is no prior knowledge of the occluded location.

Fig. 6. Sample images from the FRGC database with randomly located occlusions.

The FRGC database contains 8,014 images from 466 subjects in two sessions. These images contain variations such as illumination and expression changes, time-lapse, etc. Similar to the work in [7], an image set of 100 subjects (eight images in two sessions are selected for each subject), is used in experiments. To simulate the randomly located occlusions, we create an occluded image set by replacing a randomly located square patch (size of 10% to 50% of the original image) from each image in the original image set with a black block (Fig. 6). We design experiments according to the three occlusion scenarios: **Uvs.O**, **Ovs.U** and **Ovs.O**. There are 2,400 testing samples for each scenario. All images are cropped and re-sized to 80×65 pixels and the patch size is 6×5 pixels (the effect of patch size is discussed in Section III-D1).

1) Uvs.O: For each subject, we select $K = 1, 2, 3$ and 4 unoccluded images respectively to form the gallery sets and use the other four images with synthetic occlusions as the probe set. Fig. 7 shows the recognition results with different values of K. The correct identification rates of all methods increase when more gallery images are available (i.e., greater value of K). When there are multiple gallery images per class and no occlusion (level = 0%) in images, HMM performs better than the supervised method SVM and the local matching based NBNN. But its performance is significantly affected by the increasing occlusions. In addition, when $K = 1$, HMM performs worst among all methods since there are not enough gallery images to train a HMM for each class. For NBNN and DICW, using the difference patch achieves better results than using the original patch (i.e., OP-NBNN and OP-Warp). Especially, by comparing DICW with OP-Warp, and NBNN with OP-NBNN, it can be found that difference patches improve the results of DICW more significantly than that of NBNN. As introduced in Section II-A, the difference patches are generated by the spatially continuous patches so they enhance the *order information* within a patch sequence, which is compatible with DICW. With the optimal location weights, NBNN-ub and OP-NBNN-ub perform better than SVM. When $K = 1, 2, 3$ and 4, the average rates for the six occlusion levels of DICW are 2.3%, 4.3%, 5.5% and 4.4% better than that of NBNN-ub, respectively. When the occlusion level = 0%, the performance of SRC is better than DICW. However, the performance drops sharply when the degree of occlusion increases. When $K = 1$, the *Image-to-Class* distance degenerates to the *Image-to-Image* distance. DICW, which allows *time warping* during matching, still achieves better results while the level of occlusion increases.

2) Ovs.U and Ovs.O: We fix the value of K to 4 and consider that occlusions exist in the galley set. For each occlusion level (from 0% to 50%), we conduct experiments with the

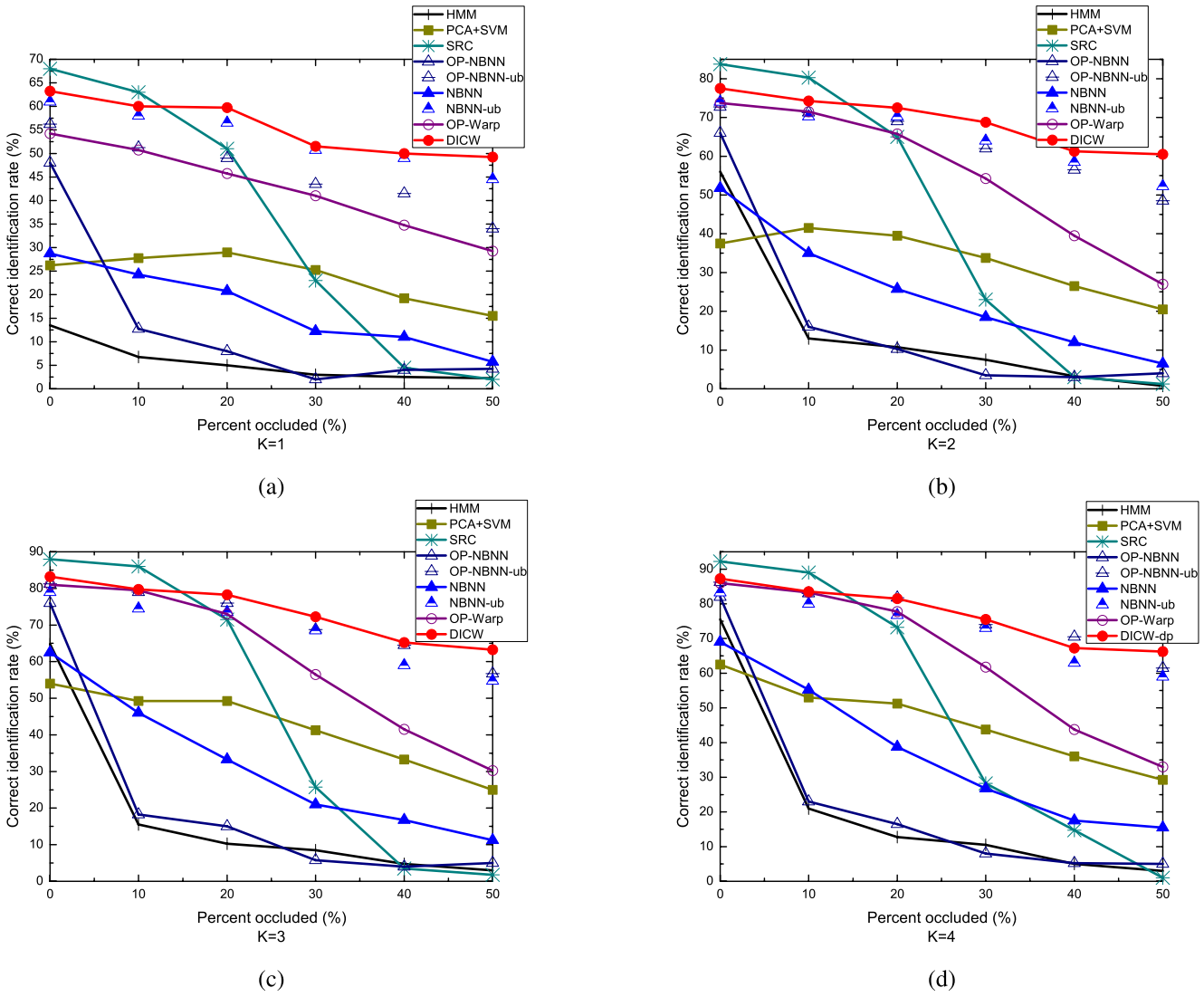


Fig. 7. **Uvs.O**: identification results on the FRGC database with different number of gallery images per subject: a) $K = 1$, b) $K = 2$, c) $K = 3$ and d) $K = 4$.

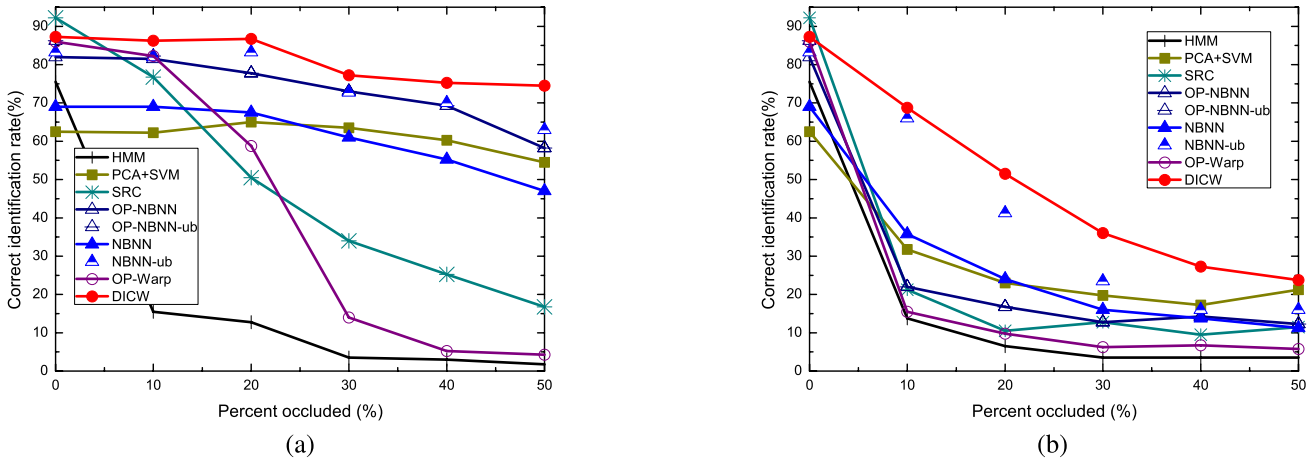


Fig. 8. a) **Ovs.U** and b) **Ovs.O**: identification results on the FRGC database with occlusions in gallery or/and probe sets.

following settings: 1) 400 occluded images (four images per subject) from the original set as the gallery set and 400 images from the unoccluded set as the probe set (**Ovs.U**) and 2) 400 occluded images as the gallery set and 400 occluded images as the probe set (**Ovs.O**). Note that the images in the gallery set are different from those in the probe sets. Fig. 8 shows the

recognition results. The methods (e.g., HMM, SVM, SRC) which include occluded gallery images for training/modelling perform poorly in these two cases. NBNN does not perform consistently in **Ovs.U** and **Ovs.O**. Using the original patch (i.e., OP-NBNN) performs better than using the difference patch (i.e., NBNN) in **Ovs.U**. For DICW, using the difference patch is always better than using the original patch (i.e., OP-Warp). This confirms that the difference patch works better with DICW, as analysed before. DICW outperforms the best of NBNN (i.e., NBNN-ub) by a larger margin of 5.5% (Fig. 8a) and 8.1% (Fig. 8b) on average than that (4.4% in Fig. 7d) in the **Uvs.O** tested with $K = 4$. These results confirm the effectiveness and robustness of DICW when the gallery and probe images are occluded. On the whole, our method performs consistently and outperforms other methods in all three occlusion cases.

B. Face Identification With Facial Disguises

We next test the proposed method on the AR database [26] which contains real occlusions. First, we consider that no occlusion is present in both gallery and probe sets. Next, we conduct experiments according to the three occlusion cases. DICW does not rely on the prior knowledge of occlusions. We will demonstrate that it works well in both general and difficult situations later.

The AR database contains over 4,000 colour images of 126 subjects' faces. For each subject, 26 images in total are taken in two sessions (two weeks apart). These images suffer from different variations in facial expressions, illumination conditions and occlusions (i.e., sunglasses and scarf, as shown in Fig. 5). Similar to the works in [6]–[8], [20], [36], and [37], a subset of the AR database (50 men and 50 women) is used [38]. All images are cropped and re-sized to 83×60 pixels and the patch size is 5×5 pixels.

1) Without Occlusion: We have evaluated the performance of DICW when no occlusion exists in both gallery and probe sets in Section III-A (i.e., occlusion level = 0% in the experiments). In this section, we adopt the setting in [8] using images without occlusions to further test DICW. For each subject, 14 images are chosen (four neutral faces with different illumination conditions and three faces with different expressions in each session). Seven images from Session 1 are used as the gallery set and the other seven from Session 2 as the probe set. Table II shows the identification rates. HMM does not perform as good as others. This may be due to other variations such as illumination and expression changes in the training images. Again, the difference patch does not improve NBNN comparing with the original patch (i.e., OP-NBNN). With the optimal location weights, the difference patch (i.e., NBNN-ub) is 3.7% better than the original patch (i.e., OP-NBNN-ub). For DICW, using the difference patch is 3.1% better than using the original patch (OP-Warp). As analysed in Section III-A, the difference patch can enhance the relative *order* of adjacent patches, the results in Table II also indicates that the difference patch is more compatible with these methods which considers the *order information*. When there is no occlusion in the gallery and probe images, both reconstruction based method (e.g., SRC) and local matching based methods (e.g., NBNN and DICW) achieve relatively satisfactory results. DICW significantly outperforms NBNN and is still slightly better than the upper bound of NBNN (i.e., NBNN-ub).

TABLE II
IDENTIFICATION RESULTS ON THE AR DATABASE WITHOUT OCCLUSIONS (K = 7)

Method	Correct identification rate (%)
HMM [35]	66.5
PCA+SVM [32]	89.7
SRC [8]	92.0
OP-NBNN [34]	89.6
OP-NBNN-ub [34]	92.0
NBNN [34]	85.3
NBNN-ub [34]	95.7
OP-Warp	93.6
Proposed DICW	**96.7**

2) Uvs.O: The unoccluded frontal view images with various expressions are used as the gallery images (eight images per subject). For each subject, we select $K = 1, 2, 4, 6$ and 8 images to form the gallery sets, respectively. Two separate image sets (200 images each) containing sunglasses (cover about 30% of the image) and scarves (cover about 50% of the image) respectively are used as probe sets. Fig. 9 shows the recognition results. The correct identification rates increase when more gallery images are available. HMM and SVM are generic training based methods and are unable to deal with *unseen* occlusions in the probe images. In the scarf testing set, the performance of SRC deteriorates significantly compared with that on the sunglasses set due to the occluded area is much larger. Local matching based NBNN and DICW perform better than others on the whole. With the optimal location weights, NBNN-ub achieves very comparable performance to DICW. But DICW is slightly superior. Even at $K = 1$, DICW still achieves 90% and 83% on the sunglasses set and scarf set, respectively.

With the same experimental setting, we also compare DICW with the state-of-the-art algorithms (using eight gallery images per subject, $K = 8$). The results are shown in Table III. Only the pixel intensity feature is used except the MLERPM method. MLERPM, which is also a local matching based method as ours, uses SIFT [39] and SURF [40] features to handle the misalignment of images. Note that compared with other methods, DICW does not require training. It achieves comparable or better recognition rates among these methods and with a relatively low computational complexity (see Section III-D6). In the scarf set, albeit the fact that nearly half of the face is occluded, only 2% images are misclassified by DICW. To the best of our knowledge, this is the best result achieved on the scarf set under the same experimental setting.

3) Ovs.U and Ovs.O: For the **Ovs.U** scenario, we select four images with sunglasses and scarves to form the gallery set and eight unoccluded images as the probe set. For the **Ovs.O** scenario, we conduct two experiments: 1) two images with

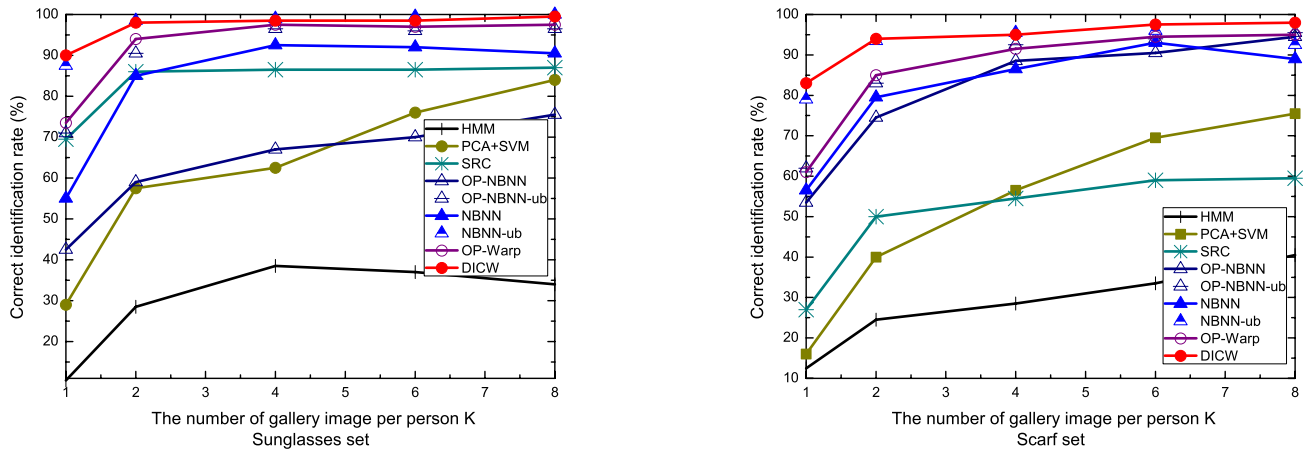


Fig. 9. **Uvs.O**: identification results on the AR database with (a) sunglasses occlusion and (b) scarf occlusion.

TABLE III

UVS.O: COMPARISON OF DICW AND THE-STATE-OF-THE-ART METHODS (K = 8)

Method	Sunglasses	Scarf	Average	Feature
SRC-partition[1] [8]	97.5	93.5	95.5	Greyscale
LRC [9]	96.0	26.0	61.0	
CRC-RLS-partition[1] [11]	91.5	95.0	93.3	
SEC-MRF [41]	99.0∼**100**	95.0∼ 97.5	97.0∼**98.8**	
l_{struct} [36]	99.5	87.5	93.5	
OP-Warp	97.5	95.0	96.3	
Proposed DICW	99.5	**98.0**	**98.8**	
MLERPM [37]	98.0	97.0	97.5	SIFT [39] & SURF [40]

[1] *SRC-partition* and *CRC-RLS-partition* indicate the strategy of partitioning an image into 4×2 patches for performance improvement for the original method SRC [8] and CRC-RLS [11], respectively.

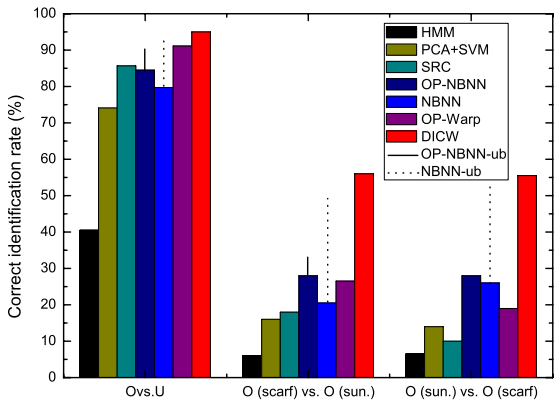


Fig. 10. **Ovs.U** and **Ovs.O**: identification results on the AR database with occlusions in gallery or/and probe sets.

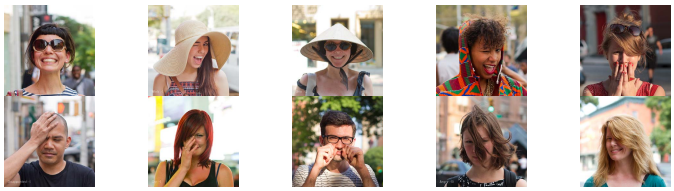


Fig. 11. Sample images from the TFWM database.

scarves as the gallery set and two images with sunglasses as the probe set; 2) vice versa. Note that with this setting, in each test the occlusion type in the gallery set is *different* from that in the probe set, which is very challenging for recognition.

The results are shown in Fig. 10. On the gallery set which contains occluded faces, the results of HMM and SVM are much worse than others as expected. In the **Ovs.O** testing, there are only two gallery images per subject. It is very difficult for SRC to reconstruct an unoccluded probe image with such limited number of gallery images. Local matching based NBNN and DICW perform better. Comparing OP-NBNN with OP-NBNN-ub, and NBNN with NBNN-ub, it can be found that the performance of NBNN is highly dependent on the optimal location weights. Overall, DICW consistently outperforms the best of NBNN (i.e., NBNN-ub) by about 4% on average.

C. Face Identification With General Occlusions in Realistic Environment

In this Section, we test our method on the *The Face We Make* (TFWM) [28] database captured under natural and arbitrary conditions. It has more than 2,000 images which contains frontal view faces of strangers on the streets with uncontrolled lighting. The sources of occlusions include glasses, sunglasses, hat, hair and hand on the face. Besides occlusions, these images also contain expression, pose and head rotation variations. In our experiments, we use images of 100 subjects

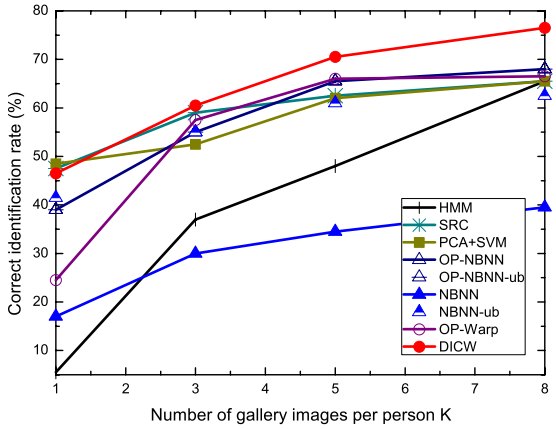


Fig. 12. Identification rates (%) on the TFWM database.

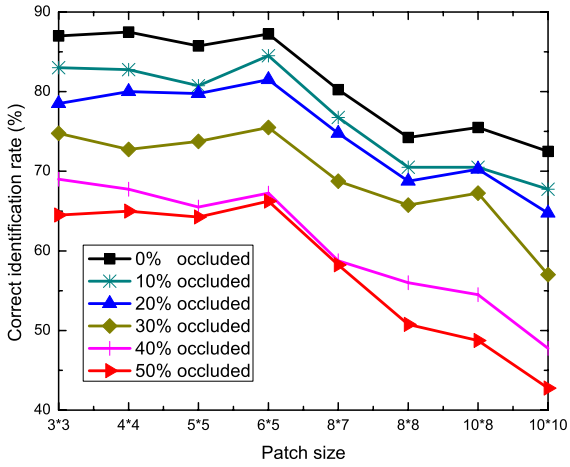


Fig. 13. Identification rates (%) with respect to the patch size.

(ten images per subject) containing various types of occlusions (Fig. 11). For each subject, we choose $K = 1, 3, 5$ and 8 unoccluded images as gallery sets, respectively, and the remaining two images as the probe set. Occlusions occur at random in the gallery or probe set or in both. This includes all the three occlusion scenarios in Section I. The face area of each image is cropped from the background and re-sized to 80×60 pixels and the patch size is 5×5 pixels. Only the pixel intensity feature is used in all methods.

The recognition results are shown in Fig. 12. Note that the images used in the experiments are *not well aligned* due to the uncontrolled variations. Some occlusions (e.g., hand) have very similar texture as the face, which are difficult to be detected by skin colour based models [42]. NBNN, which only relies on the texture similarity without considering the structural constraint of a face, does not achieve comparable performance as ours. As more gallery images are available, the accuracies of all methods increase. When $K = 8$, most methods reach a *bottleneck* with the rate around 65%. DICW outperforms these methods by a notable margin.

D. Discussion

1) The Effect of Patch Size: To investigate this the impact of patch size on the performance, we use 400 unoccluded images (size of 80×65 pixels) of 100 subjects from the FRGC database as the gallery set and 400 images in each of six probe sets, which contain randomly located occlusions from 0% to 50% level, respectively. We test DICW with the patch sizes from 3×3 pixels to 10×10 pixels. Note that we employ this dataset because the location and size of the occlusions is independent to the patch size.

The correct identification rates with respect to the patch size are shown as Fig. 13. There is no sharp fluctuation in each of the rate curve when the patch size is less than or equal to 6×5 pixels. Our method is robust to different patch sizes in an appropriate range despite the ratio of occlusions. The relatively smaller patches lead to better recognition rate since they provide more flexibility to use spatial information than the larger ones. Based on the experimental results, sizes smaller than 6×5 pixels are recommended.

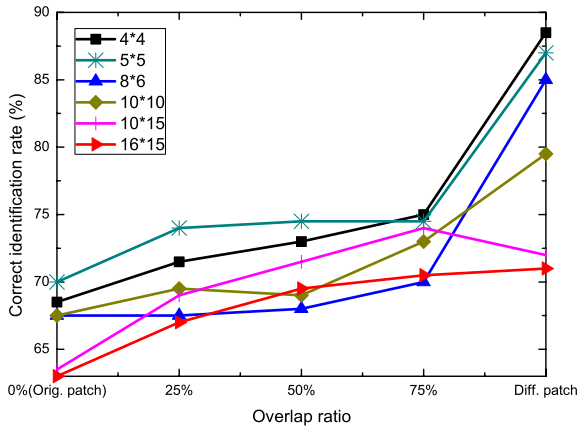


Fig. 14. Identification rates (%) with respect to the overlap ratio comparing with using the difference patches.

2) The Effect of Patch Overlap: In the previous experiments we used the difference patch to enhance the textured features in patches. It is interesting to see if the overlapping patch has this similar effect. We conducted experiments on the AR database to investigate this since it contains real occlusions with different textures. We selected four unoccluded images from Session 1 for each subject as the gallery set and two images with sunglasses and scarves from Session 2 as the probe set so the testing dataset contains variations of occlusions and illumination changes. We tested the use of different patch sizes (4×4 to 16×15 pixels) with different overlap ratios (0%, 25%, 50%, 75%) and compared their results with that of using the difference patch. 25% ratio means the adjacent patches have a 25% horizontal overlap. So the larger the ratio is, the larger the number of patches will be in each image sequence. Note that 0% overlap ratio means using the original patches (with intensity features).

Fig. 14 shows the recognition results. On the whole, large overlap ratio leads to better accuracy. Note that higher overlap ratio also increases the number of patches in each image sequence, which leads to a higher computational cost. For small patch sizes (i.e., 4×4, 5×5 and 8×6 pixels), using the difference patch yields significantly better results than using

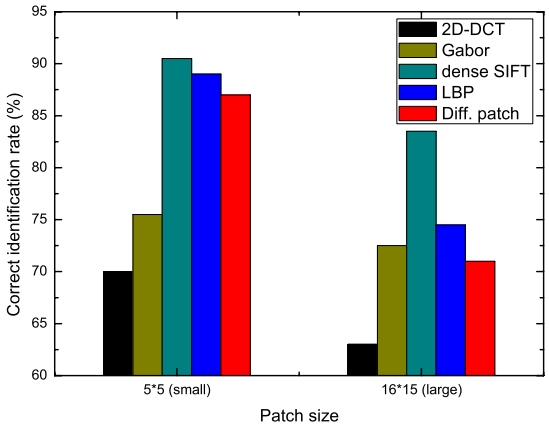


Fig. 15. Identification rates (%) of using different image descriptors and the difference patches.

the overlapping patch. This is compatible with our analysis in Section II-A. A difference patch is the approximation of the first-order derivative of adjacent *small* patches. The first-order derivative operator is sensitive to edges, which is able to enhance the textured regions. When the patch becomes large (i.e., 10×10, 10×15 and 16×15 pixels), the advantage of using the difference patch is not obvious. This is reasonable since the texture in a large patch is less uniform. Note that the overall performance of using the small patch is better than that of using the large patch. DICW is compatible with the small patch as analysed in Section III-D1 so in the experiments we used the best one, the difference patch instead of the overlapping patch.

3) The Effect of Image Descriptor: In Section III-D2, our experiments indicate that the difference patch leads to better accuracy since it is able to enhance the textured regions in a face image. In this section we will carry out experiments to compare the discriminative power of the proposed difference patches and other local image descriptors such as 2D-DCT (Discrete Cosine Transform coefficients), Gabor [43], LBP [31] and dense SIFT [44]. We use the same dataset in Section III-D2 and test both small patch size (i.e., 5×5 pixels) and large patch size (i.e., 16×15 pixels).

Fig. 15 shows the recognition results. The 2D-DCT feature is not as discriminative as others so it performs worst. For large patch size, as we analysed before, the difference patch does not perform very well. For small patches, the performance of difference patch is comparable with that of SIFT and LBP. The Gabor features do not perform better than the difference patch since the patch is too small to extract discriminative features. Note that the computation of difference patch is much simpler than other images descriptors. From Fig. 15 we can see, the local image descriptor is able to strengthen DICW when the image contains uncontrolled variations such as illumination changes and occlusions. When dealing with the uncontrolled data, applying these local features can further improve the performance of DICW.

4) Robustness to Misalignment: The face registration error can largely degrade the recognition performance [1] as we mentioned in Section I. To evaluate the robustness of DICW

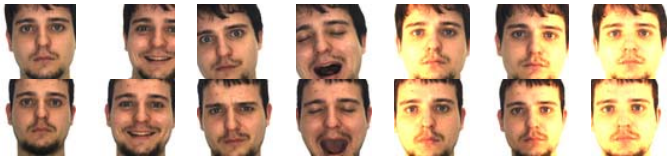


Fig. 16. Sample images of the same subject from the AR database without alignment (AR-VJ).

TABLE IV
IDENTIFICATION RATE (%) ON THE AR-VJ DATASET

Method	Correct identification rate (%)
Av-SpPCA[1][47]	93.6
DCT[1][1]	95.3
SURF-Face [48]	95.9
P2DW-FOSE [45]	**98.2**
Proposed DICW	97.3

[1] Using manually aligned images

to the misalignment of face images, we use a subset of the AR database with 110 subjects (referred to AR-VJ) used in the work in [45]. The faces in AR-VJ are automatically detected by the Viola & Jones detector [46] and cropped directly from the images without any alignment. Different from the images in the original AR database which are well cropped (Fig. 5), these images contain large crop and alignment errors as shown in Fig. 16.

Following the same experimental setting in [45], seven images of each subject from the first session are used as the gallery set and the other seven images from the second session as the probe set. All images are re-sized to 65×65 pixels and the patch size is 5×5 pixels. As analysed in the last section, we use the $\mathrm{LBP}^{u2}_{8,2}$ descriptor [31] for feature extraction to handle the illumination variations.

The recognition results are shown in Table IV. DICW outperforms other methods and achieves very close result to P2DW-FOSE [45], which is also a training-free method like ours. But different from DICW, which performs warping on the *patch level*, P2DW-FOSE is a pseudo 2D warping method on the *pixel level* and its time complexity is quadratic in the number of pixels [45].

5) Extension to Face Verification in the Wild: In this Section, we extend DICW for face verification tasks using the *Labeled Faces in the Wild* (LFW) database [29], which is the most active benchmark for FR. The task of face verification under the LFW database's protocol is to determine if a pair of face images belongs to the same subject or not. Note that in the verification of each pair, it is a *Image-to-Image* comparison. So the experiments on the LFW database can be considered as an evaluation for the effectiveness of DICW when only *time warping* is used (no *within-class warping*).

The LFW database contains 13,233 face images of 5,749 subjects collected from the Internet. These images are captured in uncontrolled environments and contain large variations in pose, illumination, expression, time-lapse and various types of occlusions (Fig. 17). Following the testing

Fig. 17. Sample images from the LFW database (six matched image pairs for six subjects).

protocol of *View 2*, we use the most difficult experimental setting: *restricted unsupervised setting* where no class label information is available. In *View 2*, there are 3,000 *matched* (i.e., positive) and 3,000 *mismatched* (i.e., negative) image pairs. They are equally divided into ten randomly generated sets and the final verification performance is evaluated by the ten-fold cross-validation. Here image pairs are classified into *the same subject* or *different subjects* by thresholding on their distance. We use the LFW-a version and $\text{LBP}^{u2}_{8,2}$ as feature descriptor. All images are cropped and re-sized to 150×80 pixels and the patch size is 3×3 pixels.

Chen *et al.*'s work [49] produces very competitive results on the LFW database by using the high-dimensional LBP feature. It is confirmed that features sampled at facial landmarks lead to better recognition performance than those sampled from regular grids. Motivated by this, we also select 25 landmarks [50] of the inner face and follow the similar process as in [49]: 1) normalise the unaligned images according to 2 facial landmarks (i.e., the tip of the nose and the centre of the mouth), and 2) extract image blocks (size of 30×30 pixels) centred around 25 facial landmarks from each image. Each block is partitioned into 3×3 pixels patches which are then concatenated to form a sequence. The original DICW algorithm is performed according to each block (i.e., sequence) and a corresponding distance is generated respectively. The sum of these distances is the final distance for each image pair. We refer our method with this strategy (i.e., sampling features around landmarks) as DICW-L and the original DICW (i.e., sampling features from regular grids) as DICW-G.

LFW is an extremely challenging database containing large uncontrolled variations, especially pose changes. As presented in [51], the first several principal components (PCs) usually capture these uncontrolled variations in the principal component analysis (PCA) subspace [33]. Therefore, we adopt the component analysis process in [51] to remove the first several PCs for performance improvement by:

$$\boldsymbol{F}' = \boldsymbol{F} - \boldsymbol{X}_i \boldsymbol{X}_i^T \boldsymbol{F} \tag{7}$$

where $\boldsymbol{F}$ is the original feature vector of an image by concatenating all the patch features of the image sequence (i.e., $\boldsymbol{P}$ or $\boldsymbol{G}_k$ in Section II-B) and $\boldsymbol{X}_i$ is the first i components in the PCA subspace. We quantitatively test the value of i using the *View 1* dataset provided by the LFW database and set the optimal value $i = 8$. $\boldsymbol{F}'$ is the improved feature vector used in the experiments for the LFW database. In this way, the large uncontrolled variations can be reduced to some extent. At the same time, different from the general dimension reduction operation (i.e., the original PCA), the topological structure of each image is still maintained so our patch based DICW can be performed directly on the improved features by this process.

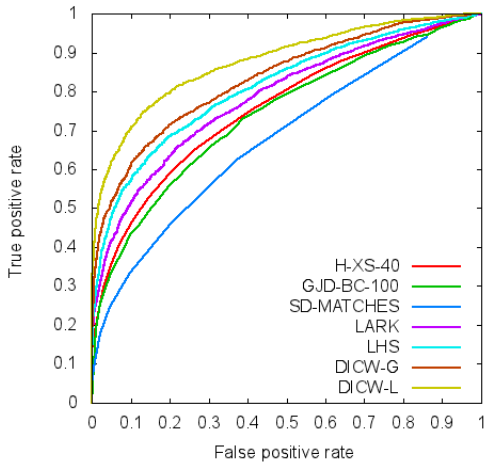


Fig. 18. ROC curves of the-state-of-the-art methods and DICW on the LFW database.

TABLE V
AREA UNDER ROC CURVE (AUC) ON THE LFW DATABASE UNDER UNSUPERVISED SETTING

Method	AUC	Feature extraction
SD-MATCHES [54]	0.5407	From grids
H-XS-40 [54]	0.7574	
GJD-BC-100 [54]	0.7392	
LARK [52]	0.7830	
LHS [53]	0.8107	
Proposed DICW-G	0.8286	
Proposed DICW-L	**0.8740**	From landmarks

We compare DICW with other methods under the same testing protocol *without outside training data*. In the experiments, only $\text{LBP}^{u2}_{8,2}$ descriptor [31] is used. We draw the the ROC (Receiver Operating Characteristic) curves of DICW and other state-of-the-art methods in Fig. 18. It shows the performance of DICW-G is better than other methods which use only single feature such as SD-MATCHES (SIFT [39]), H-XS-40 (LBP [31]), GJD-BC-100 (Gabor [43]), LARK (locally adaptive regression kernel descriptor [52]) and LHS (local higher-order statistics [53]). When extracting features around facial landmarks, the performance of DICW is further improved with a large margin. The area under the ROC curve (AUC) of DICW-L is 0.874 as shown in Table V, which is the best among all methods. These experimental results confirm the effectiveness of DICW even only *time warping* is performed.

6) Computational Complexity and Usability Analysis: From Algorithm 1 in Section II-C we can see that the time complexity of DICW for computing the distance between a query image and an enrolled class is $\boldsymbol{O}(\max\{M, N\} l K)$, where M, N are the numbers of patches in each probe sequence and gallery sequence, respectively. l is the window width as mentioned in Section II-B. For better readability, here we use M' to represent $\max\{M, N\}$. The number of gallery images per class K is very small compared with the number

TABLE VI
COMPARISON OF AVERAGE RUNTIME (S)

	Per class	All class
SRC [8]	N/A	89
Proposed DICW	0.05	6

of patches M' in each sequence (i.e., $K \ll M'$). Thus the complexity is represented as $\boldsymbol{O}(M'l)$. Note that usually $l = 10\%M'$, so the warping distance can be obtained very efficiently. On the other hand, the computational cost of the reconstruction based method SRC is very high [8]. To facilitate intuitive comparisons, Table VI shows the runtime of DICW and SRC[3] for classifying a query image under the same setting as the experiments of Table III using Matlab implementation (running on a platform with quad-core 3.10GHz CPUs and 8 GB memory). DICW is about 15 times faster than SRC [8] when classifying a query image.

Compared with the reconstruction based approaches, which represent a query image using all enrolled images, DICW computes the distance between the probe image and each enrolled class independently. So in the real FR applications, the distance matrix can be generated in parallel and the enrolled database can be updated incrementally. This is very practical for the real-world applications.

IV. FURTHER ANALYSIS AND IMPROVEMENT

In the previous sections, we evaluate DICW using extensive experiments with face images with various uncontrolled variations. In this section we will further analysis why the DICW works compared with similar methods, and when and why it will fail. We also discuss the idea for improving the performance of DICW.

NBNN [34] presented in the previous sections is a similar method to ours. It also calculates the *Image-to-Class* distance between a probe patch set and a gallery patch set from a given class. The difference is that it does not consider the spatial relationship between patches like ours and each probe patch can be matched to any patches from any location in the gallery patch set. Fig. 19 is an illustration example. The occluded probe image is from class 74 but is incorrectly classified to the class 5 by NBNN. Actually the images from class 74 and class 5 are not alike. But the texture of sunglasses is very similar to that of beard in class 5. Without the location constraint, the beard patches are wrongly matched to the sunglasses thus the distance is affected by this occlusion. On the other hand, DICW keeps the order information and matches patches within a proper range which leads to correct classification.

NBNN calculates the distance between two patch sets and the overall distance is the sum of patch-pair distances. On the other hand, in DICW, the probe and gallery patch set are ordered. The spatial relationship between patches is encoded. When a probe patch is matched to a gallery patch, the following probe patches will only be matched to the gallery patches within a proper range. This is guaranteed by the four constraints mentioned in Section II-B. In addition, with the help of *Dynamic Programming*, DICW actually tries every possible combination of matching correspondence of patch pairs so the final matching is the global optimum for the probe patch set and the gallery patch set. Compared with NBNN, DICW considers both the texture similarity and the geometry similarity of patches. The work in [23] points out that the contextual information between facial features plays an important role in recognition. Our work confirms their observation. Although a location weight can be adopted in NBNN, the weight needs to be manually set for different testing dataset as analysed before, which is not suitable for practical applications. In DICW, the order constraint is naturally encoded during distance computation.

DICW represents a face image as a patch sequence which maintains the facial order of a face. To some extent, the geometric information of a face is reduced from 2D to 1D. However, the direct 2D image warping is an NP-complete problem [55]. P2DW-FOSE mentioned in Section III-D4 is a pseudo 2D warping method but with a remarkably large computational cost (i.e., quadratic in the number of pixels) [45]. DICW incurs a lower computational cost due to its patch sequence representation. In addition, each patch still contains the local 2D information which is helpful for classification.

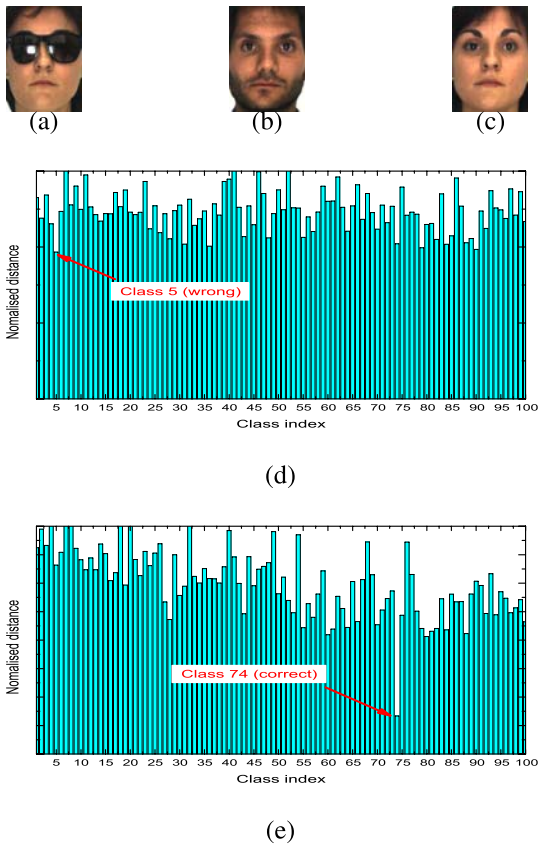


Fig. 19. (a) The probe image from class 74. (b) Classification result (class 5) by NBNN. (c) Classification result (class 74) by DICW. Distance to each class computed by (d) NBNN and by (e) DICW.

[3]We use the *l1_ls* package for implementation. http://www.stanford.edu/~boyd/l1_ls/

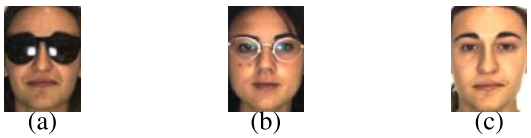


Fig. 20. (a) A probe image from class 51. (b) The wrong class (class 72) classified by DICW. (c) The gallery image from class 51.

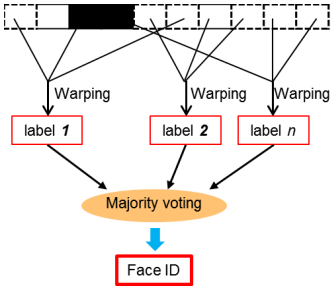


Fig. 21. Random selection and majority voting scheme for improving the performance of DICW.

A. Improving the Performance With Random Selection and Majority Voting Scheme

Fig. 20 shows a fail example which can not be correctly classified by both DICW and NBNN. The discriminative eyes region is occluded by sunglasses, which makes recognition difficult. In addition, a probe face with sunglasses (Fig. 20a) is more similar to a gallery face with glasses (Fig. 20b) in the feature space, which leads to misclassification.

Looking back to the definition of DICW in Section II-B, although *warping* is helpful for avoiding large distance error caused by occlusions, the occluded area is not directly removed during matching. Here we employ a simple but very effective scheme for improving the performance of DICW. As shown in Fig. 21, we do not use all patches in a probe sequence for warping, instead, we randomly select a subset of patch set then compute the *Image-to-Class* distance based on this subset. We repeat this n times and generate a class label (the class with the shortest distance) each time according to the calculated distance. Finally, the final class label is decided by majority voting by n *experts*. With random selection, it is possible to skip the occluded patches. It is also possible that the occluded patches are chosen but this effect will be eliminated by the majority voting strategy since we assume that the occluded areas only take up a small part of a face. This assumption is reasonable since if most parts of a face are occluded, even a human being will feel difficult to recognise it. Different from the *occlusion detection* based methods which attempt to detect and remove occlusion area as we mentioned before, this simple strategy does not rely on any prior knowledge nor any data-dependent training.

Here we use the same setting to Section III-D2. We randomly select 15% patches in a sequence each time as an *expert* and select $n = 50$ *experts* in total. Since this scheme is based on random selection, we repeat the whole classification process ten times and calculate the average identification rate. The results are shown in Table VII. The performance of DICW is improved by 2% on average by using only 50 *experts* (Note that for each *expert*, the computation of DICW is much faster than before since the number of *subset* patches is much smaller than that of the whole sequence). Generally, more experts will lead to higher accuracy since this increases the diversity of decision *views*, which is more robust to different variations. But this will also raise the whole computational cost, which needs to be considered to keep a balance between accuracy and computation. The improvement is more obvious when the number of image per class is limited. A preliminary study of using this scheme to improve DICW when $K = 1$ is discussed in [56].

TABLE VII
IDENTIFICATION RATES (%) OF DICW AND THE IMPROVEMENT SCHEME ON THE AR DATABASE

# Img./class (K)	1	2	3	4
DICW	81.0	83.5	86.0	87.0
Improvement scheme	84.5	85.2	86.5	89.0

V. CONCLUSION AND FUTURE WORK

We have addressed the problem of face recognition with occlusions in uncontrolled environments. Different from most of the current works, we consider the situation that occlusions exist in both gallery and probe sets. We proposed a novel approach, Dynamic Image-to-Class Warping (DICW), which considers the contextual order of facial components, for the recognition of occluded faces. We first represent a face image as an ordered sequence, then treat the image matching problem as the process of finding optimal alignment between a probe sequence and a set of gallery sequences. Finally, we employ the Dynamic Programming technique to compute the *Image-to-Class* distance for classification. Extensive experiments on the FRGC, AR, TFWM and LFW face databases show that DICW achieves promising performance when handling various types of occlusions. In the most challenging cases where occlusions exist in both gallery and probe sets and only a limited number of gallery images are available for each subject, DICW still performs satisfactorily. DICW can be applied directly to face images without performing occlusion detection in advance and does not require a training process. All of these make our approach more applicable in real-world scenarios. Given its merits, DICW is applicable and extendible to deal with other problems caused by local deformations in FR (e.g., the facial expression problem), as well as other object recognition problems where the geometric relationship or contextual information of features should be considered.

REFERENCES

[1] H. K. Ekenel and R. Stiefelhagen, "Why is facial occlusion a challenging problem?" in *Proc. IAPR 3rd Int. Conf. Biometrics (ICB)*, 2009, pp. 299–308.
[2] M. Storer, M. Urschler, and H. Bischof, "Occlusion detection for ICAO compliant facial photographs," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2010, pp. 122–129.
[3] D. Lin and X. Tang, "Quality-driven face occlusion detection and recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–7.

[4] T. Hosoi, S. Nagashima, K. Kobayashi, K. Ito, and T. Aoki, "Restoring occluded regions using FW-PCA for face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2012, pp. 23–30.

[5] R. Min, A. Hadid, and J.-L. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Mar. 2011, pp. 442–447.

[6] H. Jia and A. M. Martinez, "Face recognition with occlusions in the training and testing sets," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Sep. 2008, pp. 1–6.

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

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

[9] A. I. Naseem, B. R. Togneri, and C. M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010.

[10] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, vol. 6316. 2010, pp. 448–461.

[11] D. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 471–478.

[12] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.

[13] X. Wei, C.-T. Li, and Y. Hu, "Robust face recognition under varying illumination and occlusion considering structured sparsity," in *Proc. Int. Conf. Digit. Image Comput. Techn. Appl. (DICTA)*, 2012, pp. 1–7.

[14] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 748–763, Jun. 2002.

[15] X. Tan, S. Chen, Z.-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. 16, no. 4, pp. 875–886, Jul. 2005.

[16] X. Tan, S. Chen, Z.-H. Zhou, and J. Liu, "Face recognition under occlusions and variant expressions with partial similarity," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 2, pp. 217–230, Jun. 2009.

[17] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193–1205, May 2013.

[18] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Subspace learning from image gradient orientations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2454–2466, Dec. 2012.

[19] J. Zhu, D. Cao, S. Liu, Z. Lei, and S. Z. Li, "Discriminant analysis with Gabor phase for robust face recognition," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar./Apr. 2012, pp. 13–18.

[20] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2618–2625.

[21] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognit.*, vol. 39, no. 9, pp. 1725–1745, 2006.

[22] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[23] R. Wang, Z. Lei, M. Ao, and S. Z. Li, "Bayesian face recognition based on Markov random field modeling," in *Proc. IAPR 3rd Int. Conf. Biometrics (ICB)*, vol. 5558. 2009, pp. 42–51, doi: 10.1007/978-3-642-01793-3_5.

[24] X. Wei, C.-T. Li, and Y. Hu, "Face recognition with occlusion using dynamic image-to-class warping (DICW)," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.

[25] X. Wei, C.-T. Li, and Y. Hu, "Robust face recognition with occlusions in both reference and query images," in *Proc. Int. Workshop Biometrics Forensics*, Apr. 2013, pp. 1–4.

[26] A. M. Martinez and R. Benavente, "The AR face database," Autonomous Univ. Barcelona, Barcelona, Spain, Tech. Rep. 24, 1998.

[27] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 947–954.

[28] D. Miranda. *The Face We Make*. [Online]. Available: http://www.thefacewemake.org, accessed 2014.

[29] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[30] S. U. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2012, pp. 99.1–99.11.

[31] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[32] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, Apr. 2011. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[33] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1991, pp. 586–591.

[34] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[35] A. V. Nefian and M. H. Hayes, III, "Hidden Markov models for face recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 5. May 1998, pp. 2721–2724, doi: 10.1109/ICASSP.1998.678085.

[36] K. Jia, T.-H. Chan, and Y. Ma, "Robust and practical face recognition via structured sparsity," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, vol. 7575. 2012, pp. 331–344.

[37] R. Weng, J. Lu, J. Hu, G. Yang, and Y.-P. Tan, "Robust feature set matching for partial face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 601–608.

[38] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[39] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[40] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis. (ECCV)*, vol. 3951. 2006, pp. 404–417.

[41] 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)*, Sep./Oct. 2009, pp. 1050–1057.

[42] W.-H. Lai and C.-T. Li, "Skin colour-based face detection in colour images," in *Proc. IEEE Int. Conf. Adv. Video Signal-Based Surveill. (AVSS)*, Nov. 2006, pp. 56–61.

[43] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[44] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jun. 2006, pp. 2169–2178.

[45] L. Pishchulin, T. Gass, P. Dreuw, and H. Ney, "Image warping for face recognition: From local optimality towards global optimization," *Pattern Recognit.*, vol. 45, no. 9, pp. 3131–3140, 2012.

[46] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[47] K. Tan and S. Chen, "Adaptively weighted sub-pattern PCA for face recognition," *Neurocomputing*, vol. 64, pp. 505–511, Mar. 2005.

[48] P. Dreuw, P. Steingrube, H. Hanselmann, and H. Ney, "SURF-face: Face recognition under viewpoint consistency constraints," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2009, pp. 7.1–7.11.

[49] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3025–3032.

[50] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[51] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2013, pp. 348–353.

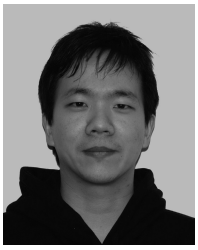
[52] H. J. Seo and P. Milanfar, "Face verification using the LARK representation," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 4, pp. 1275–1286, Dec. 2011.

[53] G. Sharma, S. ul Hussain, and F. Jurie, "Local higher-order statistics (LHS) for texture categorization and facial analysis," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, vol. 7578. 2012, pp. 1–12.
[54] J. Ruiz-del Solar, R. Verschae, and M. Correa, "Recognition of faces in unconstrained environments: A comparative study," *EURASIP J. Adv. Signal Process.*, vol. 2009, pp. 1:1–1:19, Jan. 2009.
[55] D. Keysers and W. Unger, "Elastic image matching is NP-complete," *Pattern Recognit. Lett.*, vol. 24, nos. 1–3, pp. 445–453, Jan. 2003.
[56] X. Wei and C.-T. Li, "Fixation and saccade based face recognition from single image per person with various occlusions and expressions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2013, pp. 70–75.

Xingjie Wei received the B.E. degree in computer science and technology from Central South University, Changsha, China, in 2009, and the Ph.D. degree in computer science from University of Warwick, Coventry, U.K., in 2014. She is currently a Post-Doctoral Researcher with the School of Computing Science, Newcastle University, Newcastle upon Tyne, U.K. She has authored a number of research papers in the field of face analysis and recognition. Her current research interests include biometrics, multimedia forensics, and computer vision.

Chang-Tsun Li received the B.E. degree in electrical engineering from the Chung-Cheng Institute of Technology (CCIT), National Defense University, Bade City, Taiwan, in 1987, the M.Sc. degree in computer science from the Naval Postgraduate School, Monterey, CA, USA, in 1992, and the Ph.D. degree in computer science from the University of Warwick, Coventry, U.K., in 1998. He was an Associate Professor with the Department of Electrical Engineering, CCIT, from 1998 to 2002, and a Visiting Professor with the Department of Computer Science, Naval Postgraduate School, in the second half of 2001. He is currently a Professor with the Department of Computer Science, University of Warwick. He was the Editor-in-Chief of the *International Journal of Digital Crime and Forensic* from 2009 to 2013. He is currently an Associate Editor of the EURASIP *Journal on Image and Video Processing*. He has been involved in the organization of many international conferences and workshops, and also served as a member of the International Program Committees for many international conferences. His research interests include biometrics, digital forensics, multimedia security, computer vision, image processing, pattern recognition, evolutionary computation, machine learning, data mining bioinformatics, and content-based image retrieval.

Zhen Lei received the B.S. degree in automation from the University of Science and Technology of China, Hefei, China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010, where he is currently an Associate Professor. His research interests are in computer vision, pattern recognition, image processing, and in particular, face recognition. He has authored over 80 papers in international journals and conferences. He serves as the Area Chair the of International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015.



Dong Yi received the B.S. degree in electronic engineering in 2003, the M.S. degree in communication and information system from Wuhan University, Wuhan, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009. His research areas are unconstrained face recognition, heterogeneous face recognition, and deep learning for facial analysis. He has authored and acted as a reviewer for tens of articles in international conferences and journals. He has developed face biometric algorithms and systems for the Shenzhen-Hong Kong Immigration Control Project, the 2008 Beijing Olympic Games, and the 2010 Shanghai World Expo.

Stan Z. Li received the B.Eng. degree from Hunan University, Changsha, China, the M.Eng. degree from the National University of Defense Technology, Changsha, and the Ph.D. degree from the University of Surrey, Surrey, U.K. He is currently a Professor with the National Laboratory of Pattern Recognition and the Director of the Center for Biometrics and Security Research with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, where he is also the Director of the Center for Visual Internet of Things Research. He was with Microsoft Research Asia, Beijing, from 2000 to 2004, as a Researcher. Prior to that, he was an Associate Professor with Nanyang Technological University, Singapore. He was elevated to the IEEE fellow for his contributions to the fields of face recognition, pattern recognition, and computer vision.

His research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He has authored over 200 papers in international journals and conferences, and authored and edited eight books. He was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is the Editor-in-Chief of the *Encyclopedia of Biometrics*. He served as the General Cochair of the IEEE International Conference on Automatic Face and Gesture Recognition in 2011, serves/served as the Program Cochair of the International Conference on Biometrics in 2007, 2009, and 2015, the International Joint Conference on Biometrics in 2014, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015, and has been involved in organizing other international conferences and workshops in the fields of his research interest.