ROBUST FACE RECOGNITION WITH OCCLUSIONS IN BOTH REFERENCE AND QUERY IMAGES

Xingjie Wei, Chang-Tsun Li and Yongjian Hu

Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK

{x.wei,c-t.li}@warwick.ac.uk, yongjian.hu@dcs.warwick.ac.uk

ABSTRACT

Face recognition (FR) systems in real environment need to deal with uncontrolled variations in face images such as occlusions and disguise. Most of the current FR algorithms do not consider the fact that occlusions may exist in both reference and query images. In this paper, we summarise three occlusion cases that a realistic FR system should take account of. We present a novel non-parametric classification method to handle the occlusion related problems. Our method represents a face image as a sub-patch sequence which maintains the inherent structure information of the face. Matching is based on the *Image-to-Class* distance from a query sequence to all reference sequences of an enrolled class. Experimental results on public databases verify the effectiveness and robustness of the proposed method.

Index Terms— Face recognition, Biometrics, Occlusion, Image-to-Class distance, Dynamic Time Warping

1. INTRODUCTION

Partial occlusions (e.g., sunglasses, scarf) in face images are great challenges for FR. The intra-class variations are usually larger than the inter-class variations due to occlusions, which degrades the recognition performance. A large number of approaches have been proposed for dealing with occlusion related problems. Reconstruction based methods[1, 2, 3] achieve good performance recently. An unoccluded image is recovered from an occluded probe image (i.e., query face) by a linearly combination of gallery images (i.e., enrolled faces) then the probe image is assigned to the class (i.e., person) with the minimal reconstruction error. Most of these methods assume that occlusions only exist in probe images and the gallery/training images are *clean*.

However, in the real-world environment, occlusions may occur in both reference and query data, especially in surveillance scenarios. Another example is that, for human-computer-interaction (HCI) applications which are totally open to users, reference faces may be occluded in the enrolment stage. The approaches mentioned above are easily affected by occlusions in the training data. When the number of gallery images is limited, to discard these affected images would, on the one hand, lead to *small sample size (SSS)* problem; on the other hand, lose useful information for recognition[4].

We summarise three occlusion cases in Tab.1, which a FR system may encounter in the real world. Most of the current methods rely on a clean gallery/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. Jia et al. proposed a reconstruction based method[5], as well as an improved SVM[6] to handle occlusions in the training set. But the methods depend on a preprocessing of occlusion detection through the use of skin colour. Chen et al.'s work[4] uses the low-rank matrix recovery to deal with this problem. However, it requires faces to be well registered in advance. For face images, the registration by aligning annotated landmarks (e.g., the centres of the eyes) may not be accurate due to occlusions[7].

Table 1. Three typical occlusion cases in the real world.

	Gallery	Probe	Scenarios
Uvs.O:	Unoccluded	Occluded	Access control, boarder check
Ovs.U:	Occluded	Unoccluded	Suspect detection, shoplifter recognition
Ovs.O:	Occluded	Occluded	

We take a different view to deal with occlusion problems in above scenarios. We employ the Dynamic *Image-to-Class* Warping (DICW) framework which is first introduced in our previous work[8]. An image is firstly partitioned into sub-patches which are then concatenated in the raster scan order (i.e., from left to right and top to bottom) to form a sequence. Our method does not need a training phase. Matching is implemented by calculating the *Image-to-Class* distance from a probe sequence to all the gallery sequences of an enrolled class, in which each patch in the probe sequence can be matched to a patch from different gallery sequences (Fig.1). In addition, *the order of facial features (i.e., from forehead, eyes, nose and mouth to chin) maintained in each sequence does not change despite occlusions or imprecise registration.* Taking this into account, our method uses location constraints[9] for matching patches to enforce the

978-1-4673-4989-5/13/$31.00

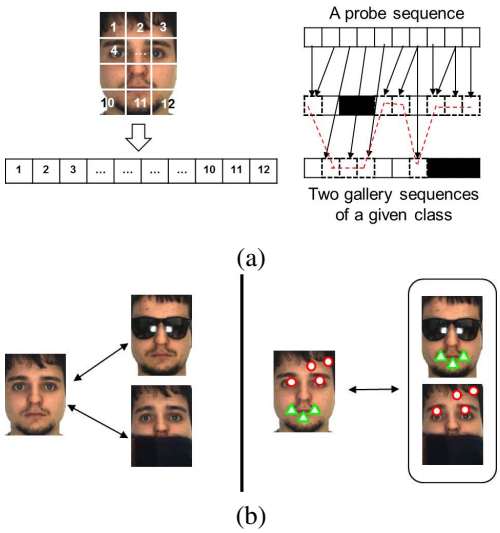


Fig. 1. (a): A face is represented as a sequence. Then the *Image-to-Class* distance is computed from a probe sequence to the gallery sequences of a given class. Black blocks indicate the occluded patches. The arrows indicate the matching correspondence between patches and the dashed line marks the optimal warping path. (b) The illustration of the *Image-to-Image* (left) and the *Image-to-Class* (right) distances. Matched features are indicated by the same symbol.

order information.

Our method tests all possible matching correspondence between patches and selects the combination with minimal overall cost for the whole sequence. So the occluded patches which cause large distance error can be ignored. Moreover, when occlusions exist in the gallery set, our model is capable of exploiting the visible information from different gallery images. Thus, our method, which uses the *Image-to-Class* distance and considers the inherent structure of the face, is able to deal with occlusions in both reference and query images. The results of comprehensive experiments on public databases verify the effectiveness of the proposed method.

2. DYNAMIC IMAGE-TO-CLASS WARPING

In our model, a probe image is represented as a sub-patch sequence as $\boldsymbol{P} = \{\boldsymbol{p}_1, ..., \boldsymbol{p}_m, ..., \boldsymbol{p}_M\}$ where $\boldsymbol{p}_m$ is the feature vector extracted at the m-th patch and M is the number of patches. We denote the gallery set of a given class containing K images as $\boldsymbol{G} = \{\boldsymbol{G}_1, ..., \boldsymbol{G}_k, ..., \boldsymbol{G}_K\}$. Each gallery image is similarly represented as a sequence of N patches $\boldsymbol{G}_k = \{\boldsymbol{g}_{k1}, ..., \boldsymbol{g}_{kn}, ..., \boldsymbol{g}_{kN}\}$.

Motivated by the time series analysis technique Dynamic Time Warping (DTW)[9], the order of facial features which included in each sequence can be viewed as the temporal order. Like in DTW, a warping path $\boldsymbol{W}$ which indicates the matching correspondence of patches between $\boldsymbol{P}$ and $\boldsymbol{G}$ of T temporal steps is defined as $\boldsymbol{W} = \{w_1, w_2, ..., w_T\}$. The t-th element w_t is an index triplets (m_t, n_t, k_t) which indicates that patch $\boldsymbol{p}_m$ is matched to patch $\boldsymbol{g}_{kn}$ at time t where $m_t \in \{1, 2, ..., M\}, n_t \in \{1, 2, ..., N\}$ and $k_t \in \{1, 2, ..., K\}$. As in DTW, the warping path also satisfies the *boundary*, *continuity* and *monotonicity* constraints[9] which maintain the temporal order. Considering the context of FR, we use a window $|m_t - n_t| \leqslant l$, where l is the window width to constrain the range for matching two patches in different positions. This is reasonable since a probe patch (e.g., eye) should not be matched to a patch (e.g., mouth) that is too far away. l is set to 10% of $max(M, N)$[9].

Unlike DTW, which computes the distance only between two time sequences, our model calculates the *Image-to-Class* distance. The distance between a probe image $\boldsymbol{P}$ and the gallery set of a given class (person) $\boldsymbol{G}$ is defined as:

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

where $C_{w_t} = C_{m_t, n_t, k_t} = \|\boldsymbol{p}_m - \boldsymbol{g}_{kn}\|_2$ is the Euclidean distance between two patches $\boldsymbol{p}_m$ and $\boldsymbol{g}_{kn}$. From Eq.(1) we can see, the optimal warping path is the path with the minimal overall cost. This optimization problem can be solved by *Dynamic Programming* (DP). We define the cumulative cost $D_{m,n,k}$ to be the cost of assigning a m-patch sequence to a set of n-patch sequences and matching the m-th patch $\boldsymbol{p}_m$ to a patch from the k-th gallery image. $D_{m,n,k}$ can be computed recursively by DP as:

$$D_{m,n,k} = \min \begin{pmatrix} \boldsymbol{D}_{\{(m-1,n-1)\}\times\{1,2,...,K\}} \\ \boldsymbol{D}_{\{(m-1,n)\}\times\{1,2,...,K\}} \\ \boldsymbol{D}_{\{(m,n-1)\}\times\{1,2,...,K\}} \end{pmatrix} + C_{m,n,k} \tag{2}$$

Computing the cumulative cost matrix is implemented using a 3-D table, indexed by (m, n, k). Thus, $DICW(\boldsymbol{P}, \boldsymbol{G})$ in Eq.(1) can be obtained by $\min_{k\in\{1,2,...,K\}} \boldsymbol{D}_{M,N,k}$. Since the probe and gallery images are usually processed to the same scale in FR systems, we set $M = N$. After computing the distances between a probe image $\boldsymbol{P}$ to all classes, $\boldsymbol{P}$ is assigned to the nearest class.

Discussion

From Eq.(2) we can see, the order information, which represents the inherent structure of a face, is considered during matching since each step is based on the set of predecessors that satisfy the warping path constraints[9]. **Different from patch-wise matching,** our method tries every possible warping path and selects the one with minimal overall cost. So the occluded patches which cause large distance error would be ignored. The *Image-to-Class* distance is the *globally* optimal cost for matching.

In addition, as shown in Fig.1a, a patch of the probe image can be matched to patches of K different gallery images. 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 a unoccluded patch at the same location is thus higher. When occlusions occur in probe and/or 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.1b).

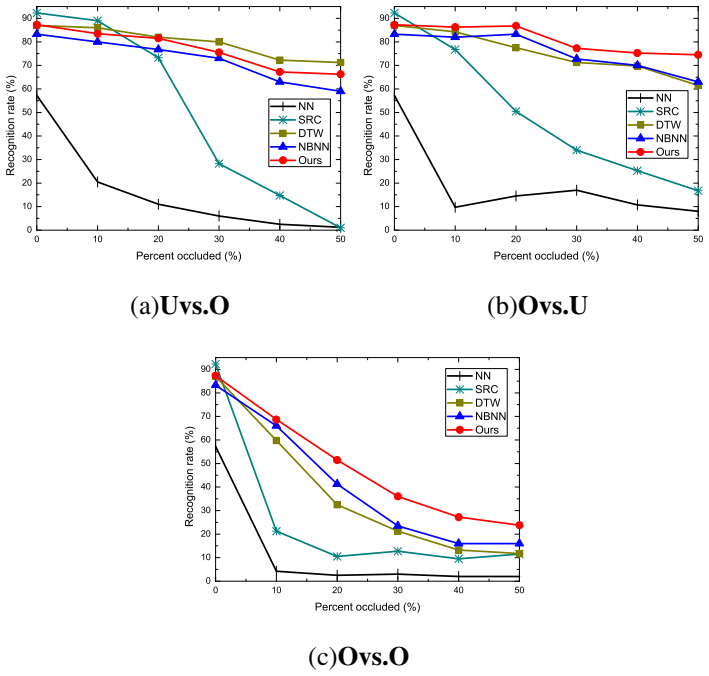
3. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, a series of experiments are conducted on three databases according to the three cases described in Sec.1. We quantitatively compare our method with four representative methods in the literature: the reconstruction based Sparse Representation Classification (SRC)[1] which achieves good performance for FR recently, the original Dynamic Time Warping (DTW)[9] which considers the order information as ours, the Naive Bayes Nearest Neighbor (NBNN)[10] which also uses the *Image-to-Class* distance as ours, and the baseline Nearest Neighbor (NN).

Similar to[8], in our method we use the grey values of difference patches as features, which are computed by applying subtraction between the grey values of two neighbouring patches. These difference patches generated by the spatially continuous patches are able to enhance the order information within a sequence. For the patch-based DTW and NBNN, we test the original patches and difference patches respectively and report the best results. The grey level images are used in SRC and NN. We test different patch sizes though extensive experiments[8] and the sizes from 4×4 to 6×5 pixels are recommended.

3.1. In the FRGC database

This section evaluates our method using images with synthetic occlusions. We select a subset containing 100 subjects from the FRGC2.0 (Face Recognition Grand Challenge) database[11]. To simulate the contiguous occlusion, we create an occluded set by replacing a randomly located square patch (size from 10% to 50% of the image) from each image in the original set with a black block. For each occlusion level (0% to 50%), we set three experiments: 1) 400 unoccluded images (4 images per person) from the original set as the gallery set and 400 images from the occluded set as the probe set **(Uvs.O)**, 2) 400 occluded images as the gallery set and 400 unoccluded images as the probe set **(O vs.U)** and 3) 400 occluded images as the gallery set and 400 occluded images as the probe set **(Ovs.O)**. Note that in each experiment, the gallery set does not intersect the probe set. The locations of occlusion blocks in the gallery and probe sets are *random* and unknown to the algorithm. All images are cropped and re-sized to 80×65 pixels. The patch size is 6×5 pixels.



(a)**Uvs.O** (b)**Ovs.U**

(c)**Ovs.O**

Fig. 2. Recognition rates on the FRGC database.

Fig.2 shows the recognition results in the three cases with different levels of occlusions. The performance of SRC is slightly better than others when the occlusion level$\leqslant$10%, however, drops sharply as the occlusion increases. As discussed before, reconstruction based methods like SRC are easily affected by occlusions in the gallery data. In the **Uvs.O** case, DTW and our method which consider the inherent structure of the face perform better than NBNN. In the following two cases, our *Image-to-Class* method outperforms others, especially when both the gallery and probe images are occluded (Fig.2c). On the whole, our method performs consistently in all three occlusion cases.

3.2. In the AR database

In this section, we evaluate our method using images with real disguises. A subset[12] of the AR database[13] (50 men and 50 women) is used in our experiments. All images are cropped and re-sized to 83×60 pixels and the patch size is 5×5 pixels. Similar to Sec.3.1, we conduct three series of experiments: for each person, we select

1. **Uvs.O**: 8 unoccluded images as the gallery set, 2 images with sunglasses and 2 images with scarves as 2 separated probe sets, respectively.
2. **Ovs.U**: 4 images with sunglasses and scarves as the gallery set and 8 unoccluded images as the probe set.
3. **Ovs.O**: 1) 2 images with scarves as the gallery set and 2 images with sunglasses as the probe set; 2) vice versa.

Fig. 3. Examples of the realistic occluded images.

Note that with this setting, in each test the occlusion type in the gallery set is *different* from that in the probe set.

The recognition results are shown in Tab.2. The performance of SRC in the **Ovs.O** case is poor since the task is very challenging. The sunglasses cover about 30% of the face and the scarves covers nearly 50%. Unlike DTW and NBNN whose recognition rates fluctuate according to different occlusions (from sunglasses to scarf), our method behaves consistently and outperforms other methods in most cases.

Table 2. Recognition rates (%) on the AR database

		NN	SRC	DTW	NBNN	Ours
Uvs.O: Gallery-unoccluded						
Probe	sunglasses	69.5	87.0	99.0	96.5	**99.5**
	scarf	11.5	59.5	96.5	95.5	**98.0**
Ovs.U: Gallery-occluded						
Probe	unoccluded	42.6	85.7	87.7	94.4	**94.6**
Ovs.O: Gallery-occluded						
Probe	sunglasses	5.5	18.0	55.0	49.0	**56.0**
	scarf	5.5	10.0	**61.5**	52.5	55.5

3.3. In the realistic database

We also evaluate our method using a realistic database collected by Dexter Miranda[14]. It contains frontal view faces of strangers on the streets. We select a subset of this database containing 80 subjects with different types of occlusions such as sunglasses, hat, hair or hand in front of a face, as shown in Fig.3. For each person, we choose 8 images as the gallery set and the remaining 2 images as the probe set. Occlusions occur at random in the gallery or probe set or both. The face area of each image is cropped from the background and re-sized to 80×60 pixels. The patch size is 5×5 pixels.

The recognition results are shown in Tab.3. The performance of all methods is relatively poor. Note that the images used in this experiment are taken under outdoor environment with uncontrolled illumination. These images are **not well aligned** due to the various occlusions. Some occlusions (e.g., hand) have very similar texture as the face, which are difficult to be detected by skin colour based models. In this situation, SRC, DTW and NBNN perform as bad as the baseline NN. However, the proposed method still achieves relatively better performance.

4. CONCLUSION

In this paper, we presented a novel non-parametric classification method for occluded face recognition. Compared with the current methods, our method is able to deal with occlusions which exist in both gallery and probe sets. When the gallery images are contaminated, the *Image-to-Class* distance leads to better performance than the *Image-to-Image* distance. Experimental results demonstrate that the proposed method outperforms the state-of-the-art methods for occlude FR.

5. REFERENCES

[1] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Trans. PAMI*, vol. 31, no. 2, pp. 210 –227, feb. 2009.

[2] Zihan Zhou, A. Wagner, H. Mobahi, J. Wright, and Yi Ma, "Face recognition with contiguous occlusion using markov random fields," in *Proc. of ICCV 2009*, 29 2009-oct. 2 2009, pp. 1050 –1057.

[3] Meng Yang and Lei Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *Proc. of ECCV 2010*, Berlin, Heidelberg, 2010, ECCV'10, pp. 448–461, Springer-Verlag.

[4] Chih-Fan Chen, Chia-Po Wei, and Y.-C.F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. of CVPR 2012*, june 2012, pp. 2618 –2625.

[5] Hongjun Jia and A.M. Martinez, "Face recognition with occlusions in the training and testing sets," in *Proc. of FG 2008*, sept. 2008, pp. 1 –6.

[6] Hongjun Jia and A.M. Martinez, "Support vector machines in face recognition with occlusions," in *Proc. of CVPR 2009*, june 2009, pp. 136 –141.

[7] A.M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans.PAMI*, vol. 24, no. 6, pp. 748 –763, jun 2002.

[8] Xingjie Wei, Chang-Tsun. Li, and Yongjian Hu, "Face recognition with occlusion using dynamic image-to-class warping (dicw)," in *Proc. of FG 2013*, apr. 2013.

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

[10] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. of CVPR 2008*, june 2008, pp. 1 –8.

[11] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. of CVPR 2005*, june 2005, vol. 1, pp. 947 – 954 vol. 1.

[12] A.M. Martinez and A.C. Kak, "Pca versus lda," *IEEE Trans. PAMI*, vol. 23, no. 2, pp. 228 –233, feb 2001.

[13] AM Martınez and R. Benavente, "The ar face database," *CVC Tech. Rep.*, vol. 24, 1998.

[14] Dexter Miranda, "The face we make," www.thefacewemake.org.

Table 3. Recognition rates (%) using the realistic images

NN	SRC	DTW	NBNN	Ours
62.5	69.4	64.4	68.1	**75.6**