IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Combined training strategy for low-resolution face recognition with limited application-specific data

ISSN 1751-9659
Received on 29th June 2018
Revised 13th March 2019
Accepted on 11th June 2019
doi: 10.1049/iet-ipr.2018.5732
www.ietdl.org

Dan Zeng[1,2], Luuk Spreeuwers[2], Raymond Veldhuis[2], Qijun Zhao[1] ✉

[1]*College of Computer Science, Sichuan University, Chengdu, 610065, People's Republic of China*
[2]*Faculty of EEMCS, University of Twente, Enschede, The Netherlands*
✉ *E-mail: qjzhao@scu.edu.cn*

Abstract: Application-specific data, typical for certain biometric applications, are often not sufficiently available. The authors present a solution for face recognition with limited application-specific data. Existing methods often use a 'convolutional neural network (CNN) classifier' architecture with the CNNs serving as feature extractors. The CNNs are trained with massive general (i.e. not application specific) data and the classifier is trained with application-specific data. Instead, the authors propose a combined training strategy to train the classifier on a balanced mixture of general and application-specific data. The balance is achieved such that the recognition performance is maximised. The proposed method largely alleviates the needs for application-specific data. To prove its effectiveness, they apply the proposed method to low-resolution face recognition. Specifically, they use the heterogeneous joint Bayesian (HJB) classifier that is capable of comparing features from the same modality but with different characteristics. To further boost performance, the authors augment the training data of HJB by pre-processing the data to resemble application-specific data. They conducted extensive experiments on challenging data sets, namely, SCface and COX. The results show that the proposed method improves the true match rate on SCface at a false match rate of 10% by ~11% and the true match rate on COX at a false match rate of 1% by ~12%.

1 Introduction

Deep neural networks (DNNs) [1] have become popular in face recognition and yield excellent performance in many cases, especially for faces in the wild. However, DNNs require vast amounts of manually labelled training data. The publicly available general data sets like Webface [2] somehow satisfy these needs. However, in many real-world applications [e.g. low-resolution (LR) face recognition, near-infrared face recognition] the publicly available manually labelled application-specific data are too limited to train the deep networks from scratch.

Focusing on the LR face recognition problem and inspired by recent works [2–5], which use a combination of a convolutional neural network (CNN) and a traditional classifier to mitigate the needs for application-specific data, we employ the CNN joint Bayesian architecture in this paper. In particular, the architecture consisting of two identical CNNs and one heterogeneous joint Bayesian (HJB) classifier is proposed. It uses the CNN to extract features from face images and the HJB classifier to recognise the faces based on their extracted features. The HJB is capable of comparing features that are extracted from images of the same modality but in different domains, e.g. high resolution (HR) versus LR face images, visible versus near-infrared face images. We train the CNN on the general (i.e. not application-specific) data, and the HJB on the available small quantity of application-specific data. This is feasible, because having fewer parameters, the HJB needs less data for training.

Despite the superb accuracy achieved by such a hybrid architecture of DNNs and traditional classifiers, existing methods [2–5] use either general data or data from a specific application domain to train the traditional classifiers. To better exploit all the available data, we propose to train the classifier (i.e. the HJB in this paper) on a balanced mixture of general and application-specific data (see Fig. 1). The balance between application-specific and general data is chosen such that the best recognition accuracy is achieved. To the best of our knowledge, this is the first paper about exploring mixed sources of training data that are substantially different in their nature for deep learning-based face recognition. This strategy undoubtedly provides a new perspective to other applications (not limited to face recognition applications) that suffer from insufficient application-specific data.

To prove the effectiveness of such a combined training strategy, we demonstrate that the thus trained CNN-HJB outperforms the state-of-the-art of face recognition in surveillance applications (See Fig. 2). Facial images captured by surveillance cameras are usually of poor quality, particularly because of their LR, which seriously deteriorates face recognition performance and makes it difficult to locate the facial landmarks, which are needed for registration, accurately. Owing to this, we propose a combination of two identical CNNs with a heterogeneous classifier that is optimised to compare features derived from HR reference images, e.g. mugshots, with those from surveillance images. This also gives us reason to augment the training data of the HJB by preprocessed general data that resembles the application-specific data. Moreover, we use matching score based registration (MSBR) [6] to mitigate the inaccurate facial landmarking problem during operation.

The four main contributions of this paper are described below

- Combined training strategy (see Fig. 1) is proposed to largely alleviate the needs for application-specific data. It trains the classifier with a mixture of general and application-specific data, and the mixture is balanced such that the recognition performance is optimised.
- We extend the work in [7] based on HJB by implementing CNN-HJB architecture and apply it to LR face recognition. The architecture consists of two identical CNNs and a HJB classifier that can compare features extracted from images of the same modality but in different domains.
- Regarding the surveillance application, general data are preprocessed in order to resemble application-specific data to augment the training data of the HJB. Moreover, MSBR is employed to mitigate the inaccurate facial landmarking problem during operation.
- The combination of contributions (1)–(3) leads to an improvement of the state-of-the-art performance in LR face recognition as shown by our comprehensive experiments on SCface and COX databases.

© The Institution of Engineering and Technology 2019

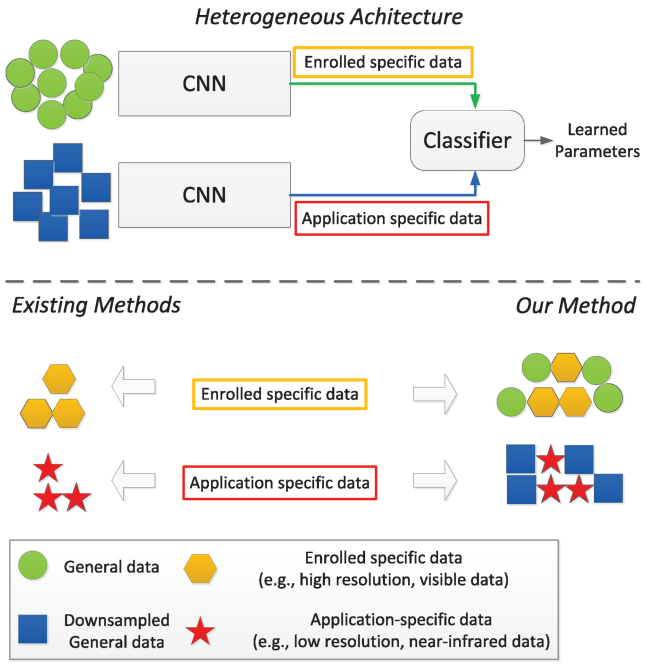

Fig. 1 *Existing methods mainly exploit application-specific data for classifier training. These methods based on heterogeneous structure still suffer from the limited application-specific data, whereas our proposed method utilises the combined training strategy by mixing the general data and application-specific data in a balanced way. In heterogeneous architecture, CNNs are treated as feature extractors to the classifier. There is no necessary for the two CNNs to be different here*

Fig. 2 *LR Challenges in Surveillance Applications: LR has been a long-standing problem in face recognition in surveillance scenarios. From left to right, these figures show LR, HR, and downsampled HR (ds HR) images. 'LR' images are captured from surveillance cameras. 'HR' images are taken under controlled conditions. 'ds HR' images are downsampled from HR images to the same resolution as 'LR'. Obvious differences can be observed in appearance between the three types of images*

The remainder of the paper is organised as follows. In Section 2, we review related work on learning with limited application-specific training data and LR face recognition. In Section 3, we present our combined training strategy together with our hybrid CNN-HJB architecture and then we apply the proposed method to LR face recognition. In Section 4, we evaluate the performance of the proposed method and compare it with the state-of-the-art methods. Moreover, we investigate the performance of the proposed method as a function of the spatial extent of the facial images and the inter-pupillary distance (IPD). Finally, in Section 5, we draw conclusions.

2 Related work

2.1 Learning with limited application-specific data

As explained before, deep learning techniques require a large amount of manually labelled training data to achieve super performance, while for many real-world applications (e.g. LR face recognition, near infrared face recognition) it is impossible to obtain massive manually labelled data or there is only limited application-specific data publicly available. Taking the surveillance applications as a case, the only LR data sets that are currently available include SCface [8], COX [9], MBGC [10], ChokePoint [11], and UCCS [12].

Nowadays few-shot learning [13] and transfer learning [14] gradually gain much attention from researchers because they are specifically designed to address the limited application-specific data problems. Hermans *et al.* [15] modify the triplet loss to generate more triplets per batch for person reidentification. Our method is different in the sense that it exploits available general data to mitigate the insufficient data problem. Feature-based transfer learning, where the features of the source task are transformed to closely match those of the target task, or a common latent feature space is discovered, is commonly used. For example, the hybrid architectures of CNN + classifier [2–5] use the CNN as a feature extractor and a traditional classifier, such as a joint Bayesian classifier or an SVM, to perform recognition in the CNN-induced feature space. Existing hybrid architectures use either part of testing data or general data to train traditional classifiers. To the best of our knowledge, this paper is the first to explore mixed sources of training data that are substantially different in their nature for deep learning.

2.2 LR face recognition

In [16] three approaches are described to LR face recognition. One is just to ignore the additional information in the HR image, and down-sample it to an LR image; the second is super resolution (SR), converting the LR image into an HR image; and the third one is resolution robust.

SR approaches have been developed over recent decades. Compared with traditional vision-oriented SR methods, recognition-oriented SR methods convert images to higher resolution while optimising discriminative properties at the same time. S^2R^2 [17] is the first work that realises SR and recognition simultaneously. Zou *et al.* [18] use a piecewise linear regression model to learn a relationship between LR and HR image spaces and then apply it to convert the LR image to an HR image. The model takes data and discriminative constraints into account. However, it is time consuming and needs many training samples from the application domain.

Resolution-robust approaches [16] treat LR face recognition as a mixed-resolution problem. Ren *et al.* [19] map LR and HR samples onto various Hilbert spaces and project them onto the learnt subspaces for comparison. The projections are learnt by minimising the dissimilarities captured by kernel Gram matrices in the LR and HR spaces. Mixed-resolution biometric comparison [20] considers the combined statistics of various resolution images and uses the likelihood ratio to compare images across resolutions. Biswas *et al.* [21] transform LR and HR images in common feature space for comparison. Transformation is learnt by iterative optimisation in such a manner that the distances between LR and HR images in the transformed space approximate those of LR images captured under the same conditions as HR images. The authors of [21, 22] address the problem of pose variations in LR images and use tensor analysis to locate facial landmarks for feature extraction. Furthermore, Mudunuri and Biswas [23] propose stereo matching to compare the two images in the transformed space [21]. Then they present an efficient reference-based method to reduce the computational cost of stereo matching without significantly affecting the recognition accuracy.

© The Institution of Engineering and Technology 2019

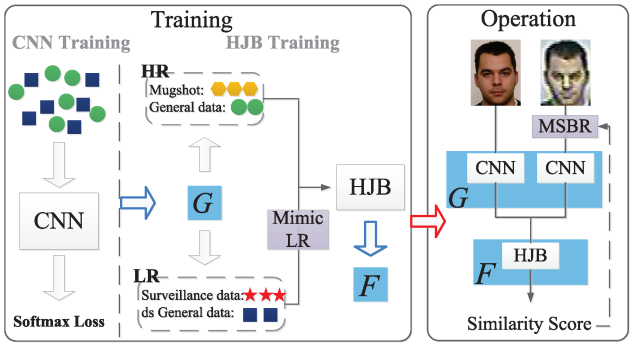

Fig. 3 *Block diagram of training and operation for LR face recognition in surveillance applications. Two identical CNNs are used during operation and both are trained on a mixture of general images of different resolutions, termed 'General data'. First, a CNN is used as a feature extractor G. Then, combined training strategy is proposed, which trains the HJB classifier with the features extracted from both general data and application-specific surveillance data, resulting in a transformation function F. 'ds General data' refers to downsampled 'General data'. Surveillance data are application-specific from the surveillance scenario. Mugshot refers to the enrolled HR faces. 'Mimic LR' means generating LR images from 'ds General data' to augment the training set for HJB. During operation, MSBR is proposed to improve the LR data registration*

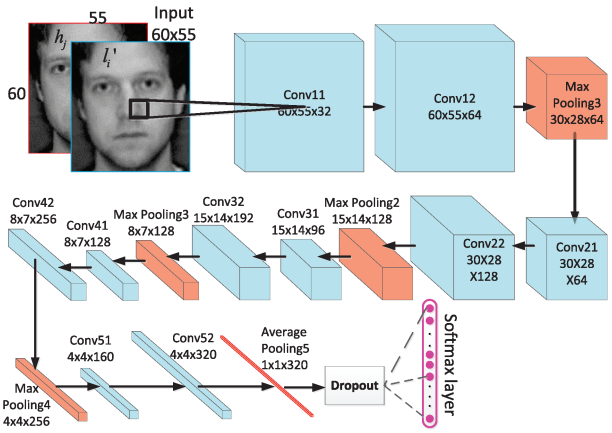

Fig. 4 *Architecture of RIDN [25] for resolution-robust feature extraction*

Table 1 Composition of RIDN data set. The table presents the number of subjects (# Sub), the number of images per subject (# Ips), and the average IPD

Data set	# Sub	# Ips	IPD, pixel
WebFace [2]	10,069	1–534	58(5)
FERET [26]	1195	1–24	60(2)
CAS-PEAL [27]	1040	3–43	61(3)
FRGC v2 [28]	466	1–88	126(2)
Multi-PIE [29]	337	83–486	72(5)
MUCT [30]	176	7–12	89(6)
Faces94 [31]	153	7–20	48(4)
AR [32]	100	2–6	57(3)
PIE [33]	68	2–5	80(8)
ORL [34]	40	6–10	34(3)
Pointing 04 [35]	15	32–42	53(5)
Grimace [36]	12	2–20	51(5)

Recently, there are two works [24, 25] using data augmentation and investigating whether CNNs are suitable for LR face recognition. Both generate LR data by downsampling HR data for augmentation. In [24], a robust partially coupled network taking into account SR, domain adaptation and robust regression, is designed to address the very LR recognition problem. In [25] it is demonstrated that training on HR data that are downsampled to a variety of resolutions can improve the recognition performance of a CNN on mixed-resolution input images. The proposed CNN, termed as resolution-invariant deep network (RIDN), is the basis of the method proposed here. RIDN performs well in LR face recognition, achieving 74% rank1 face identification rate on SCface data set following the same protocol in [23]. However, when the IPD value of the LR images drops <10 pixels, its performance decreases rapidly.

3 Proposed method

The proposed CNN-HJB framework with combined training strategy is shown in Fig. 3. Let $\boldsymbol{h} \in \mathbb{R}^M$ be a reference image which has been organised into a column vector of length M, $\boldsymbol{l} \in \mathbb{R}^m$ be a probe image which has been organised into a column vector of length m. G and F the projection functions obtained via the CNN and the HJB, respectively. G is first applied on $\boldsymbol{h}$ and $\boldsymbol{l}$. Then the similarity score s between $\boldsymbol{h}$ and $\boldsymbol{l}$ is given by

$$s = F(G(\boldsymbol{h}), G(f_{\text{sr}}(\boldsymbol{l}))), \tag{1}$$

where $f_{\text{sr}}: \mathbb{R}^m \to \mathbb{R}^M$ serves as a SR operator that transforms the probe image to the same size as the reference image. In this work, we use bicubic interpolation to implement f_{sr}. The function G projects both the galley image and the transformed probe image to the feature space learnt by the CNN. The features of the gallery and probe images are compared using the HJB, represented by the function F, to generate a similarity score based on the log likelihood ratio. Next, we first derive the computation of F, which will be presented in (5), where $\boldsymbol{x} = G(\boldsymbol{h})$ and $\boldsymbol{y} = G(f_{\text{sr}}(\boldsymbol{l}))$.

3.1 CNN-HJB

Let us first elaborate our proposed hybrid CNN-HJB architecture. CNN is derived from the RIDN [25]. It tries to extract features that are robust to varying resolutions by mixing images of different resolutions during training. The HJB classifier directly operates on image pairs of different characteristics to produce a similarity score. Here we propose a combination of the RIDN and the HJB which determines a more compact feature space for comparison with respect to real LR facial images and in return, it leads to improved recognition performance.

Fig. 4 [25] shows the detailed architecture of the employed RIDN network. It contains ten convolutional layers, every two of which are arranged in pairs. Max pooling layers follow every pair except the last pair, which is followed by an average pooling layer. The network is closed by a fully connected layer and a softmax layer, which indicates identity classes. Rectified linear units (ReLUs) are used for hidden neurons because ReLUs have better fitting abilities and can help produce highly non-linear and sparse features. The output feature map of the Pooling5 layer is taken as the deep feature representation whose dimensionality is $n = 320$.

The network is trained in a multi-class face identification task. Table 1 summarises the used training data. To generate LR images for training, we downsample the general data by various factors and then upsample them back. The upsampling operation is intended to ensure sufficiently large spatial supports for the convolutions in the CNN as well as to facilitate feature extraction. Several resolutions are mixed in training data since we observe that including more resolutions improves the performance.

HJB is applied to the deep features extracted by RIDN. Given two deep feature vectors $\boldsymbol{x} = G(\boldsymbol{h}) \in \mathbb{R}^n$ and $\boldsymbol{y} = G(f_{\text{sr}}(\boldsymbol{l})) \in \mathbb{R}^n$, we look for support for hypothesis H_{s} (the features originate from the same subject) versus hypothesis H_{d} (the features originate from different subjects). The decision that provides a maximum verification rate at a given false-acceptance rate follows from thresholding the likelihood ratio

$$\text{lr}(\boldsymbol{x}, \boldsymbol{y}) = \frac{p((\boldsymbol{x}/\boldsymbol{y})|H_{\text{s}})}{p((\boldsymbol{x}/\boldsymbol{y})|H_{\text{d}})}. \tag{2}$$

© The Institution of Engineering and Technology 2019

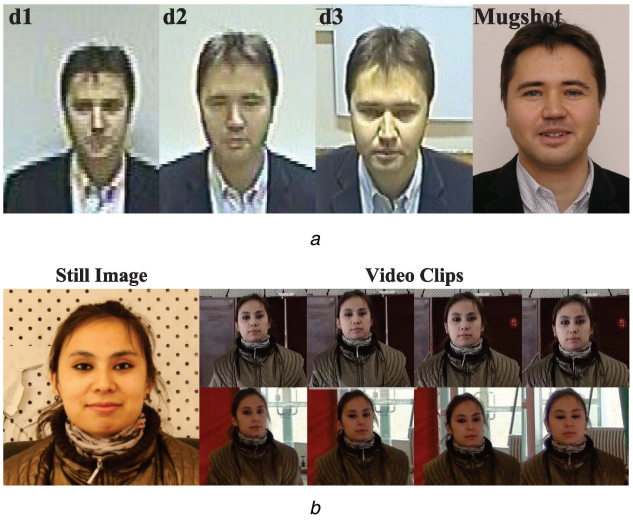

Fig. 5 *Examples of facial images in*
(a) SCface and, ***(b)*** COX data sets. In SCface, in addition to Mugshot images, surveillance images are captured at various distances of *d*1 (4.20 m), *d*2 (2.60 m), *d*3 (1.00 *m*), and Mugshot. COX contains per subject a still image (left column) of HR and video clips (right column) of LRs from various camcorders

We assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ have zero mean normal probability densities. The covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$ are $\boldsymbol{\Sigma}_{\boldsymbol{xx}} = \mathrm{E}\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$ and $\boldsymbol{\Sigma}_{\boldsymbol{yy}} = \mathrm{E}\{\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\}$, respectively. The cross-covariance matrices are $\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \mathrm{E}\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\}$ and $\boldsymbol{\Sigma}_{\boldsymbol{yx}} = \boldsymbol{\Sigma}_{\boldsymbol{xy}}^{\mathrm{T}}$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are from different individuals, then $\boldsymbol{\Sigma}_{\boldsymbol{xy}} = 0$ and $\boldsymbol{\Sigma}_{\boldsymbol{yx}} = 0$. The probability densities of the pairs of deep feature vectors are thus

$$\mathrm{p}\left(\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}\middle| H_{\mathrm{s}}\right) \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{xx}} & \boldsymbol{\Sigma}_{\boldsymbol{xy}}\\ \boldsymbol{\Sigma}_{\boldsymbol{yx}} & \boldsymbol{\Sigma}_{\boldsymbol{yy}}\end{pmatrix}\right) \tag{3}$$

$$\mathrm{p}\left(\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}\middle| H_{\mathrm{d}}\right) \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{xx}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\boldsymbol{yy}}\end{pmatrix}\right). \tag{4}$$

We estimate covariance and cross-covariance matrices during the training process. By substituting (3) and (4) into (2), taking the log and ignoring some constants, we arrive at the following similarity score

$$\begin{aligned} s &= F(G(\boldsymbol{h}), G(f_{\mathrm{sr}}(\boldsymbol{l}))) = F(\boldsymbol{x}, \boldsymbol{y})\\ &= (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}^{\mathrm{T}})\left(\begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{xx}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\boldsymbol{yy}}\end{pmatrix}^{-1} - \begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{xx}} & \boldsymbol{\Sigma}_{\boldsymbol{xy}}\\ \boldsymbol{\Sigma}_{\boldsymbol{yx}} & \boldsymbol{\Sigma}_{\boldsymbol{yy}}\end{pmatrix}^{-1}\right)\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}. \end{aligned} \tag{5}$$

This score increases monotonically with the log likelihood ratio. In order to simplify (5) and to assure that the estimated covariance matrices have full rank and can be inverted, we reduce the feature dimensionality as follows.

First we apply whitening transforms to $\boldsymbol{x}$ and $\boldsymbol{y}$, resulting in $\boldsymbol{x}_{\mathrm{w}} = \boldsymbol{W}_{\mathrm{H}}\boldsymbol{x} \in \mathbb{R}^{\mathrm{w}}$, $\boldsymbol{y}_{\mathrm{w}} = \boldsymbol{W}_{\mathrm{L}}\boldsymbol{y} \in \mathbb{R}^{\mathrm{w}}$, where $\boldsymbol{x}_{\mathrm{w}}, \boldsymbol{y}_{\mathrm{w}} \in \mathbb{R}^{\mathrm{w}}$ with dimensionality w and usually $w < n$. Thus the covariance matrices can be transformed to $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{\mathrm{w}} = \mathrm{E}\{\boldsymbol{x}_{\mathrm{w}}\boldsymbol{x}_{\mathrm{w}}^{\mathrm{T}}\} = \boldsymbol{I}$, $\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{\mathrm{w}} = E\{\boldsymbol{y}_{\mathrm{w}}\boldsymbol{y}_{\mathrm{w}}^{\mathrm{T}}\} = \boldsymbol{I}$, and $\boldsymbol{\Sigma}_{\boldsymbol{xy}}^{\mathrm{w}} = \boldsymbol{W}_{\mathrm{H}}\boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{W}_{\mathrm{L}}^{\mathrm{T}}$, where $\boldsymbol{I}$ denotes the identity matrix. We then apply singular value decomposition to obtain $\boldsymbol{\Sigma}_{\boldsymbol{xy}}^{\mathrm{w}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V} \in \mathbb{R}^{\mathrm{w}\times\mathrm{w}}$.

A compact feature dimension d is chosen so that $\boldsymbol{x}_{\mathrm{c}} = (\boldsymbol{U}_{*,1\mathrm{d}})^{\mathrm{T}}\boldsymbol{x}_{\mathrm{w}} \in \mathbb{R}^{d}$, $\boldsymbol{y}_{c} = (\boldsymbol{V}_{*,1d})^{\mathrm{T}}\boldsymbol{y}_{w} \in \mathbb{R}^{d}$, where subscript $*, 1d$ denotes only first d columns of the matrix are taken and subscript c indicates that features are mapped to a more compact common space. This way deep features are transformed from dimensionality w to d. The similarity score then becomes

$$s = (\boldsymbol{x}_{\mathrm{c}}^{\mathrm{T}}\boldsymbol{y}_{\mathrm{c}}^{\mathrm{T}})\left(\begin{pmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{pmatrix}^{-1} - \begin{pmatrix}\boldsymbol{I} & \boldsymbol{D}\\ \boldsymbol{D} & \boldsymbol{I}\end{pmatrix}^{-1}\right)\begin{pmatrix}\boldsymbol{x}_{\mathrm{c}}\\ \boldsymbol{y}_{\mathrm{c}}\end{pmatrix}, \tag{6}$$

where $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ as a diagonal matrix and d is the number of singular values ν_i of $\boldsymbol{\Sigma}_{\boldsymbol{xy}}^{\mathrm{w}}$ on the diagonal matrix. After some calculations, we finally obtain

$$s = -\sum_{i=1}^{d}\frac{\nu_i}{1-\nu_i}(\boldsymbol{x}_{\mathrm{c},i} - \boldsymbol{y}_{\mathrm{c},i})^2 + \sum_{i=1}^{d}\frac{\nu_i}{1+\nu_i}(\boldsymbol{x}_{\mathrm{c},i} + \boldsymbol{y}_{\mathrm{c},i})^2. \tag{7}$$

3.2 Combined training strategy

A typical feature-based transfer learning that can mitigate the needs for application-specific data and meanwhile achieve good accuracy is suitable for the CNN-HJB architecture. We train the HJB classifier, followed by the CNNs, with a mixture of application-specific and general data, while optimising the mixture for recognition performance. Yet, experimental results (see Section 4.5) demonstrate that performance increases if more application-specific data are used for training. However, as we remarked earlier, the availability of this type of data remains limited. A compromise approach is to consider general data as an alternative. However, obvious differences can be observed between general data and application-specific data, e.g. HR and LR face images (see Fig. 2). Simply combining these data to train a classifier is likely to the performance. Fortunately, the results in RIDN [25] show that the deep network can learn a satisfactory common feature space in which the negative effect of varying resolutions is eliminated to some extent. Thus, we exploit all the available data in different domains and believe that the combined training strategy can boost the performance when the application-specific data is very limited. In our experiments, we will further show the balance between different types of training data is also very important for good face recognition performance.

3.3 Data augmentation for HJB

We propose mimicking LR images to augment the training data for HJB. To improve the recognition accuracy at LRs, we extend the training data of the HJB by preprocessing higher resolution surveillance data to make them resemble data of LRs. For example, we use images of SCface captured at *d*2 to mimic LR data at *d*1 (see Fig. 5*a*) via downsampling with an appropriate filter. Based on the observation that edges of LR images are more highlighted compared to those of the downsampled ones (see Fig. 2), we apply a filter that preserves the mean but sharpens the edges. In this paper, a 3×3 convolution filter [−0.5,0,−0.5;0,3,0;−0.5,0,−0.5] is used.

3.4 Matching-Score based registration

Proper registration of real LR probes at runtime is difficult because automatic facial landmarking methods that are needed for registration are usually inaccurate on these images. MSBR [6, 37, 38] is thus employed to improve the robustness of the proposed method to inaccurately detected facial landmarks. For an HR reference image $\boldsymbol{h}$ and an LR probe image $\boldsymbol{l}$, given the eye-coordinates of the probe image ρ, the similarity score is written as $F(G(\boldsymbol{h}), G(\boldsymbol{l}(\boldsymbol{r}))$ (here f_{sr} is ommited for conciseness). MSBR tries to find the eye coordinates ρ^* that maximise the similarity between $\boldsymbol{h}$ and $\boldsymbol{l}(\rho)$ by registering probe images according to eye-coordinates, i.e.

$$\rho^* = \arg\max_{\rho}\ F(G(\boldsymbol{h}), G(\boldsymbol{l}(\rho))). \tag{8}$$

The final similarity score of CNN-HJB is $F^*(G(\boldsymbol{h}), G(\boldsymbol{l}(\rho^*))$. As we show in the experiments, such MSBR score calculation gains importance if the resolution of the facial images gets really very low.

4 Experiments

In this section, we first design the experiments to verify the effectiveness of different strategies in baseline RIDN-HJB architecture. Then we evaluate the performance of the proposed

© The Institution of Engineering and Technology 2019

Table 2 Effect on verification rate (VR) of different strategies in baseline RIDN-HJB.

Q10

Comb. Train	MSBR	MimicLR	VR(STD), %
N	*N*	*N*	77(4)
N	*N*	*Y*	78(4)
N	*Y*	*N*	78(4)
N	*Y*	*Y*	80(3)
Y	*N*	*N*	80(3)
Y	*N*	*Y*	81(4)
Y	*Y*	*N*	82(4)
Y	*Y*	*Y*	**84(4)**

Comb. Train, combined training of the HJB; MimicLR, mimicking LR images; STD, standard deviation.

method and compare it with state-of-the-art methods. Finally, we explore the performance of the proposed method on facial images of different IPDs.

4.1 Data sets

The RIDN data set [25] contains images chosen from 12 public face data sets. It contains 13,671 subjects, giving 438,139 images in total. Table 1 lists details of the data. The facial images display illumination, expression, and pose variations. Only facial images with poses <30° in the yaw orientation and 15° in the pitch orientation are included. All images are of relatively HR with their IPD values >20 pixels.

The SCface data set [8] contains facial images of 130 subjects taken in an uncontrolled indoor environment. Example images are shown in Fig. 5*a*. The facial images are captured by five surveillance cameras at three distances, *d*1 (4.20 m), *d*2 (2.60 m), and *d*3 (1.00 m), and one frontal mugshot per subject was taken by a digital camera is also included. The surveillance cameras are placed slightly above the subject's head. Some of the collected images are blurred. Moreover, pose and lighting as well as quality varies for different cameras at different distances. Facial images captured at *d*1 (4.20 m) are of the poorest quality compared to the other two distances. The IPD value of these images is <10 pixels.

The COX data set [9] consists of 1000 subjects, each of whom has an HR still image and several uncontrolled LR video clips. Still images of cooperative users are collected under controlled conditions and video clips are captured by various camcorders (see Fig. 5*b*). The data set is widely used in video-based face recognition such as Still-to-Video (S2V) applications.

Training data for HJB are classified into three types. In the first experiment, only the general RIDN data set is used for training and no surveillance data are used. For the second experiment, training data include surveillance data from SCface or COX. For the last experiment, combined training mixes general and surveillance data to train HJB.

4.2 Experimental settings

In our experiments, all facial images are preprocessed through face detection and facial landmarking [39], and aligned by applying affine transformation using five landmarks, i.e., left eye centre, right eye centre, nose tip, left mouth corner and right mouth corner. For all data sets mentioned above, HR and LR facial images are cropped to 60×55 and 30×24, respectively.

The CNN is trained following the method in [25], except that we mix five resolutions instead of four as used in [25]. The resolution of 55×50 is added and the increased number of resolutions results in improved performance. The experiments including CNN training are repeated five times with various random initialisations. Mean and standard deviation of the results are presented for comparison.

For HJB training, experiments on SCface are run 50 times with 100 randomly selected subjects of reference/probe combinations for training. Mean and standard deviation of the results are reported. As for COX, we follow the standard protocol that is prescribed in [9]. The finally transformed dimensionalities d (see (7)) for SCface and COX are set to 60 and 50, respectively. For combined training, we take WebFace as the general data. The subjects in WebFace are randomly selected and the number of images per subject is set to 15 and 25 for SCface and COX, respectively.

4.3 Baseline architecture RIDN-HJB

The RIDN is trained on the RIDN data set and the HJB is trained on SCface or COX. This is a special case of combined training data when the fraction of general training data for HJB equals 0. Baseline performance when different strategies are employed is given in Table 2.

4.4 Combined training strategy of the HJB

The effectiveness of the combined training is demonstrated in Fig. 6 and Table 2. Impact of the mixture of general and surveillance data on the recognition performance is also shown in Fig. 6. The horizontal axis indicates the fraction of surveillance data in the training set. The smaller the ratio is, the more general data are used. If the ratio is 1, no general training data are used. If the ratio is 0, only general training data are used. The vertical axis shows the performance in terms of verification rate at FAR = 0.1 or 0.01, depending on the data set used for evaluation.

Fig. 6*a* shows results of combined training of the HJB in SCface. For each curve, the amount of SCface data is held constant between 0.4K to 1.0K (K denotes a thousand) and the amount of general data varies. As can be seen, performance increases if more surveillance data are available for training; meanwhile, less significant performance improvement is obtained by adding only general data for training. If there is too little surveillance data, adding general data makes sense.

In Figs. 6*b* and *c*, similar observations can be made. Combined training results in an improvement in LR face recognition performance in COX. Figs. 6*b* and *c* clearly show that (i) a balanced mixture of general and application-specific data ensures the best face recognition accuracy and (ii) contribution of the proposed combined training strategy is more significant when fewer application-specific data are available for training. When the amount of COX data used for training decreases from 7.5K to 2.5K, the performance gain obtained by combined training increases from 3% to over 20%.

4.5 Mimicking LR images and MSBR

In this experiment, the variations of the eye coordinates for probe images are within [–2,2] pixels with a step of one pixel. Thus one probe image would produce 25 images of varying alignments. To reduce computational expense, only five aligned images are randomly generated based on varied coordinates. For mimicking LR, a 3×3 convolution filter [–0.5, 0,−0.5;0, 3, 0; −0.5,0,−0.5] is applied to the downsampled images to make them more similar to the surveillance data. The results are shown in Fig. 7 and Table 2, from which we can easily see that all strategies contribute to the improvement of LR face recognition performance. Combined training results in the largest improvement.

4.6 Comparison with state-of-the-art methods

Results on SCface: In this experiment, facial images of SCface captured at *d*1 (4.20 m) are compared to mugshots. The best performance reported in the literature is of RIDN [25] and MRC [7]. The comparison results of face identification and verification protocols are listed in Table 3, which prove the superiority of the proposed method.

Results on COX: We use the data set for Still-to-Still, treating each video frame as a probe for comparison which is harder than S2V. Rank-1 recognition accuracy for identification is shown in Table 4.

To make the performance of face verification comparable, we choose the video clips which appear in both versions. Video2 and video4 are selected because they correspond to videos captured by

© The Institution of Engineering and Technology 2019

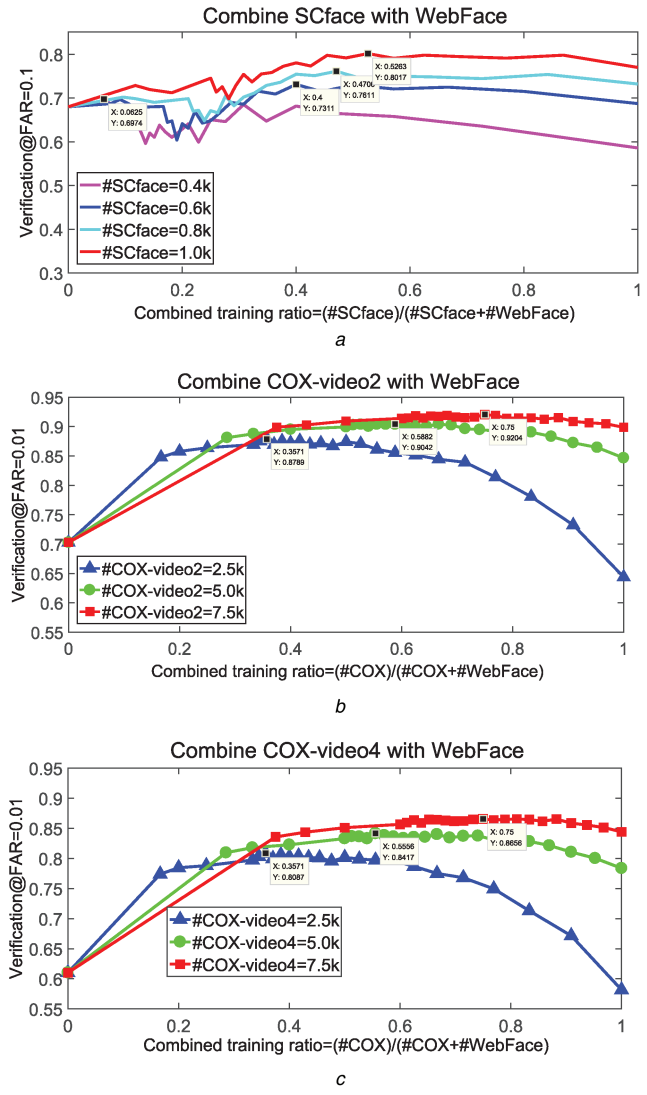

Fig. 6 *Results of combined training of the HJB in* ***(a)*** SCface, ***(b), (c)*** COX. It is shown how the mixture influences the recognition performance. When the ratio equals 0, only general data are used for HJB training. When the ratio equals 1, no general data but only surveillance data are employed (Best viewed in colour)

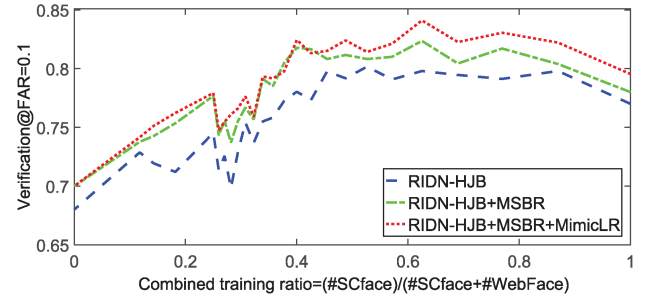

Fig. 7 *Effect on verification rate of MSBR and MimicLR for various combined training ratios. Scface data set is used in this experiment*

Table 3 Face recognition results (%) on the SCface data set

Method	VR@FAR = 0.1	Rank-1 accuracy
RIDN [25]	70(3)	28(2)
MRC [7]	73(6)	48(5)
RIDN-HJB	**84(4)**	**57(5)**

cam3 and cam2 in new version [9]. To avoid confusion, we use 'S-V2' and 'S-V3' to represent video4 and video2, respectively, in face identification and verification protocols. Face verification performance is reported in Table 5. These results again demonstrate the effectiveness of our proposed method.

4.7 Performance under different IPDs

Here we investigate the performance of the proposed method on facial images of different IPD (see Fig. 8). The results are shown in Fig. 9. As can be seen, when the IPD value changes from 10 to 6, the performance degrades slightly. However, when the IPD value drops to four the performance decreases drastically.

4.8 Discussion

Concerning the combined training strategy, it adapts to small surveillance data very well while yielding large performance improvements. Combining general data for training results in a higher performance gain if fewer LR surveillance data are available for training. MSBR gains more importance if the resolution of the facial images is lower, i.e. when finding facial landmarks becomes harder. The same holds for the mimicking of LR data by downsampling and filtering higher resolution images for training. All these contributions together result in substantial performance gains in both verification and identification settings.

We use RIDN as our CNN structure because it is designed to treat HR discriminative information and LR discriminative information equally which benefits the performance. RIDN [25] proves that considering LR information is necessary. Moreover, HJB is selected rather than SVM or other traditional classifiers because it can compare samples with different characteristics, which serve the purpose of dealing with recognition of heterogeneous images.

5 Conclusion

In this paper, we have proposed a combined training strategy for training a CNN joint Bayesian classifier with limited application-specific data and applied it to LR face recognition. RIDN and HJB classifier are taken as examples of CNN and Joint Bayesian classifier, respectively. RIDN was trained on general data whereas the HJB classifier was trained using the combined training strategy which makes it possible to exploit available general data to mitigate the insufficient data problem. More specifically, the combined training strategy mixed the general data and application-specific data in a balanced way. The balance is achieved such that the recognition performance is maximised. The results on SCface and COX show that the contribution of the combined training strategy is especially significant when a few application-specific data are available for training. For example, when the amount of COX data used for training decreases from 7.5K to 2.5K, the performance gain obtained by combined training increases from 3% to over 20%. To the best of our knowledge, this is the first paper about exploring mixed sources of training data that are substantially different in their nature for deep learning based face recognition.

Regarding the surveillance application, we mimicked LR images by preprocessing images of higher resolution with an appropriate sharpening filter to further augment the HJB training data. Moreover, matching-score based registration was employed to improve the robustness of the proposed method to poor registration of LR probes. Experiments on the challenging SCface and COX data sets have demonstrated that the proposed method outperforms state-of-the-art methods by a large margin, improving the true match rate on SCface at a false match rate of 10% by ∼11% and the true match rate on COX at a false match rate of 1% by ∼12%.

6 Acknowledgments

This work has been supported by The Royal Netherlands Academy of Arts and Sciences (KNAW) under project number 530-6CDC09.

© The Institution of Engineering and Technology 2019

Table 4 Rank-1 recognition accuracy (%) on the COX data set

Method	S-V2	S-V3
LERM [40]	42.8(1.9)	58.4(3.3)
GFK [41]	43.0(2.2)	69.8(1.7)
RIDN [25]	58.0	65.1
RIDN-HJB	**64.0(1.1)**	**71.5(0.7)**

Table 5 Face verification results (%) on the COX data set.

Method	VR@FAR = 0.01 S-V2	VR@FAR = 0.01 S-V3
LERM-ES [9]	68.37	80.46
RIDN [25]	74.41	80.56
RIDN-HJB	**86.6(0.7)**	**92.0(0.8)**

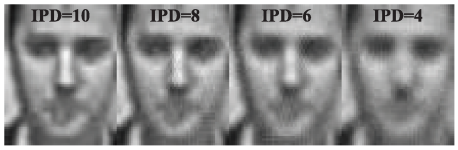

Fig. 8 *Examples of facial images in SCface data set under different IPDs. All images are rescaled to the same size for display*

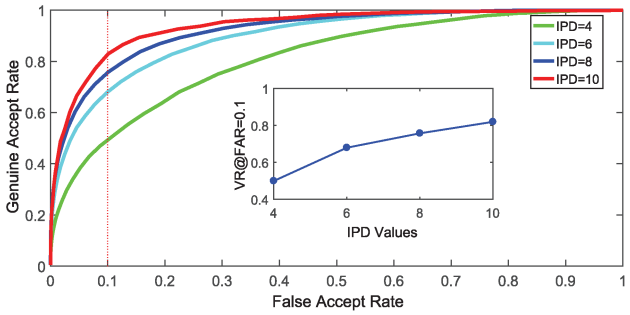

Fig. 9 *Face recognition results under different IPDs on SCface*

7 References

[1] Krizhevsky, A., Sutskever, I., Hinton, G. E.: 'Imagenet classification with deep convolutional neural networks'. Proc. Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, USA, December 2012, pp. 1097–1105

[2] Yi, D., Lei, Z., Liao, S., *et al.*: 'Learning face representation from scratch', arXiv:1411.7923, 2014

[3] Taigman, Y., Yang, M., Ranzato, M.A., *et al.*: 'Deepface: closing the gap to human-level performance in face verification'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Columbus, Ohio, USA, June 2014, pp. 1701–1708

[4] Sun, Y., Wang, X., Tang, X.: 'Deep learning face representation from predicting 10,000 classes'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Columbus, Ohio, USA, June 2014, pp. 1891–1898

[5] Sun, Y., Chen, Y., Wang, X., *et al.*: 'Deep learning face representation by joint identification-verification'. Proc. Advances in neural information processing systems (NIPS), Montreal, Quebec, Canada, December 2014, pp. 1988–1996

[6] Boom, B.J., Spreeuwers, L.J., Veldhuis, R.N.J.: 'Subspace-based holistic registration for low-resolution facial images', *EURASIP J. Adv. Signal Process.*, 2010, **2010**, (1), p. 3644

[7] Peng, Y., Spreeuwers, L., Veldhuis, R.: 'Low-resolution face alignment and recognition using mixed-resolution classifiers', *IET Biometrics*, 2017, **6**, (6), pp. 418–428

[8] Grgic, M., Delac, K., Grgic, S.: 'SCface–surveillance cameras face database', *Multimedia Tools Appl.*, 2011, **51**, (3), pp. 863–879

[9] Huang, Z., Shan, S., Wang, R., *et al.*: 'A benchmark and comparative study of video-based face recognition on COX face database', *IEEE Trans. Image Process.*, 2015, **24**, (12), pp. 5967–5981

[10] Phillips, P.J., Flynn, P.J., Beveridge, J.R., *et al.*: 'Overview of the multiple biometrics grand challenge'. Proc. Int. Conf. on Biometrics, Berlin, Heidelberg, 2009, pp. 705–714

[11] Wong, Y., Chen, S., Mau, S., *et al.*: 'Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Colorado Springs, CO, USA, June 2011, pp. 74–81

[12] Sapkota, A., Boult, T.E.: 'Large scale unconstrained open set face database'. Proc. IEEE Conf. on Biometrics Theory, Applications and Systems (BTAS), Washington, DC, USA, September 2013, pp. 1–8

[13] Ravi, S., Larochelle, H.: 'Optimization as a model for few-shot learning', 2016

[14] Afridi, M.J., Ross, A., Shapiro, E. M.: 'On automated source selection for transfer learning in convolutional neural networks', *Pattern Recognit.*, 2018, **73**, pp. 65–75

[15] Hermans, A., Beyer, L., Leibe, B.: 'In defense of the triplet loss for person re-identification', arXiv: 1703.07737

[16] Wang, Z., Miao, Z., Wu, Q.J., *et al.*: 'Low-resolution face recognition: a review', *Vis. Comput.: Int. J. Comput. Graph.*, 2014, **30**, (4), pp. 359–386

[17] Hennings-Yeomans, P.H., Baker, S., Kumar, B.V.: 'Simultaneous super-resolution and feature extraction for recognition of low-resolution faces'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Anchorage, Alaska, USA, June 2008, pp. 1–8

[18] Zou, W.W.W., Yuen, P.C.: 'Very low resolution face recognition problem', *IEEE Trans. Image Process.*, 2012, **21**, (1), pp. 327–340

[19] Ren, C.X., Dai, D.Q., Yan, H.: 'Coupled kernel embedding for low-resolution face image recognition', *IEEE Trans. Image Process.*, 2012, **21**, (8), pp. 3770–3783

[20] Peng, Y., Spreeuwers, L.J., Veldhuis, R.N.: 'Likelihood ratio based mixed resolution facial comparison'. Proc. Int. Workshop on Biometrics and Forensics (IWBF), Gjovik, Norway, March 2015, pp. 1–5

[21] Biswas, S., Bowyer, K.W., Flynn, P.J.: 'Multidimensional scaling for matching low-resolution face images', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (10), pp. 2019–2030

[22] Biswas, S., Aggarwal, G., Flynn, P.J., *et al.*: 'Pose-robust recognition of low-resolution face images', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (12), pp. 3037–3049

[23] Mudunuri, S.P., Biswas, S.: 'Low resolution face recognition across variations in pose and illumination', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, (5), pp. 1034–1040

[24] Wang, Z., Chang, S., Yang, Y., *et al.*: 'Studying very low resolution recognition using deep networks'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016, pp. 4792–4800

[25] Zeng, D., Chen, H., Zhao, Q.: 'Towards resolution invariant face recognition in uncontrolled scenarios'. Proc. IEEE Conf. on Biometrics (ICB), Halmstad, Sweden, June 2016, pp. 1–8

[26] Phillips, P.J., Moon, H., Rizvi, S., *et al.*: 'The FERET evaluation methodology for face-recognition algorithms', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (10), pp. 1090–1104

[27] Gao, W., Cao, B., Shan, S., *et al.*: 'The CAS-PEAL large-scale Chinese face database and baseline evaluations', *IEEE Trans. Syst., Man, Cybern.- A: Syst. Humans*, 2008, **38**, (1), pp. 149–161

[28] Phillips, P.J., Flynn, P.J., Scruggs, T., *et al.*: 'Overview of the face recognition grand challenge'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, June 2005, pp. 947–954

[29] Gross, R., Matthews, I., Cohn, J., *et al.*: 'Multi-PIE', *Image Vis. Comput.*, 2010, **28**, (5), pp. 807–813

[30] Milborrow, S., Morkel, J., Nicolls, F.: 'The MUCT landmarked face database', *Pattern Recog. Assoc. South Africa*, 2010, **201**

[31] 'Faces94', available at: http://cswww.essex.ac.uk/mv/allfaces/faces94.html, accessed February 2007

[32] Martinez, A.M.: 'The AR face database', CVC technical report, 1998

[33] Sim, T., Baker, S., Bsat, M.: 'The CMU pose, illumination, and expression (PIE) database'. Proc. 5th Int. Automatic Face and Gesture Recognition (FGR), Washington, DC, USA, May 2002, pp. 53–58

[34] 'ORL', available at: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html, accessed 2002

[35] Gourier, N., LetessSier, J.: 'The pointing 04 data sets'. Proc. ICPR Int. Workshop on Visual Observation of Deictic Gestures, Cambridge, UK, August 2004, pp. 1–4

[36] 'Grimace', available at: http://cswww.essex.ac.uk/mv/allfaces/grimace.html, accessed February 2007

[37] Boom, B.J., Beumer, G.M., Spreeuwers, L.J., *et al.*: 'The effect of image resolution on the performance of a face recognition system'. Proc. 9th Int. Control, Automation, Robotics and Vision (ICARCV), Singapore, December 2006, pp. 1–6

[38] Spreeuwers, L.J., Boom, B.J.: 'Better than best: matching score based face registration', *Werkgemeenschap voor Informatie-en Communicatietheorie*, 2007

[39] Sun, Y., Wang, X., Tang, X.: 'Deep convolutional network cascade for facial point detection'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Portland, Oregon, USA, June 2013, pp. 3476–3483

[40] Huang, Z., Wang, R., Shan, S., *et al.*: 'Learning Euclidean-to-riemannian metric for point-to-setet classification'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Columbus, Ohio, USA, June 2014, pp. 1677–1684

[41] Zhu, Y., Zheng, Z., Li, Y., *et al.*: 'Still to video face recognition using a heterogeneous matching approach'. Proc. IEEE Conf. on Biometrics Theory, Applications and Systems (BTAS), Washington, D.C., USA, September 2015, pp. 1–6

© The Institution of Engineering and Technology 2019

Author Queries

Q	Please make sure the supplied images are correct for both online (colour) and print (black and white). If changes are required please supply corrected source files along with any other corrections needed for the paper.
Q1	Please reduce the number of words in the Abstract to 200 words.
Q2	IET style for matrices and vectors is to use bold italics. Please check that we have identified all instances.
Q3	As per journal style references are renumbered in the text and reference list. Please confirm.
Q4	DOI information is not available in crossref.org for Ref. [38].
Q5	The initials of the author "Veldhuis R.N.J." has been changed as per the crossref.org in Ref. [6]. Please check.
Q6	The initials of the last two author have been changed as per the crossref.org in Ref. [7]. Please check.
Q7	The initials of the author "Zou w.w.w." has been changed as per the crossref.org in Ref. [18]. Please check.
Q8	Please provide page range in Ref. [30].
Q9	Please provide volume number, page range in Ref. [38].
Q10	Please provide the significance of [bold] in Tables [2-5].

© The Institution of Engineering and Technology 2019