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 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.
In Section 4, we evaluate the performance of specific training data and LR face recognition. In Section 3, we proposed method as a function of the spatial extent of the facial CNN-HJB architecture and then we apply the proposed method to images 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 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 transferred 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.
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 termed ‘General data’. 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

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

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 \( h \in \mathbb{R}^M \) be a reference image which has been organised into a column vector of length \( M \). \( l \in \mathbb{R}^m \) 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 \( h \) and \( l \). Then the similarity score \( s \) between \( h \) and \( l \) is given by

\[
s = F(G(h), G(l)) = F(G(h), G(f_du(l))).
\]

where \( f_du: \mathbb{R}^m \rightarrow \mathbb{R}^M \) 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 \( f_du \). The function \( F \) projects both the gallery 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 \( x = G(h) \) and \( y = G(f_du(l)) \).

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. 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.

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 \( x = G(h) \in \mathbb{R}^m \) and \( y = G(f_du(l)) \in \mathbb{R}^m \), we look for support for hypothesis \( H_0 \) (the features originate from the same subject) versus hypothesis \( H_1 \) (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

\[
\ln(x, y) = \frac{p(x|x|/p(x|y|/H_0)}{p(x|y|/H_0)}.
\]
where $\mathbf{D} \in \mathbb{R}^{d \times d}$ as a diagonal matrix and $d$ is the number of singular values $\nu_i$ of $\Sigma_{yy}$ on the diagonal matrix. After some calculations, we finally obtain
\begin{equation}
    s = - \sum_{i=1}^{d} \frac{y_i}{1 - \nu_i} (x_i - y_i)^2 - \sum_{i=1}^{d} \frac{y_i}{1 + \nu_i} (x_i + y_i)^2. \quad (7)
\end{equation}

### 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 perform poorly. 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 d2 to mimic LR data at d1 (see Fig. 2) 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 $3 \times 3$ convolution filter $[-0.5,0,-0.5;0,3,0;0-0.5]$ 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 $h$ and an LR probe image $l$, given the eye-coordinates of the probe image $p$, the similarity score is written as $F(G(h), G(l(p)))$ (here $f_{xy}$ is omitted for conciseness). MSBR tries to find the eye coordinates $p^*$ that maximise the similarity between $h$ and $l(p)$ by registering probe images according to eye-coordinates, i.e.
\begin{equation}
    p^* = \arg \max_p F(G(h), G(l(p))). \quad (8)
\end{equation}

The final similarity score of CNN-HJB is $F^*(G(h), G(l(p^*)))$. 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...
Table 2: Effect on verification rate (VR) of different strategies in baseline RIDN-HJB.

<table>
<thead>
<tr>
<th>Comb. Train</th>
<th>MSBR</th>
<th>MimicLR</th>
<th>VR(STD), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
<td>77(4)</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>78(4)</td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>78(4)</td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>80(3)</td>
</tr>
<tr>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>80(3)</td>
</tr>
<tr>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>81(4)</td>
</tr>
<tr>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>82(4)</td>
</tr>
<tr>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>84(4)</td>
</tr>
</tbody>
</table>

The combination of all strategies contributes to the largest improvement of LR face recognition.

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. 5a. The facial images are captured by five surveillance cameras at three distances, d1 (4.20 m), d2 (2.60 m), and d3 (1.00 m), and one frontal mugshot per subject was taken by a digital camera. 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 d1 (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 cameras (see Fig. 5b). 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. 6a 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. 6b and c, similar observations can be made. Combined training results in an improvement in LR face recognition performance in COX. Figs. 6b 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], 3, 0, -0.5, 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 d1 (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 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 530-6CD09.
Table 4  Rank-1 recognition accuracy (%) on the COX data set

<table>
<thead>
<tr>
<th>Method</th>
<th>S-V2</th>
<th>S-V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>LERM [40]</td>
<td>42.8(1.9)</td>
<td>58.4(3.3)</td>
</tr>
<tr>
<td>GFK [41]</td>
<td>43.0(2.2)</td>
<td>69.8(1.7)</td>
</tr>
<tr>
<td>RIDN [25]</td>
<td>58.0</td>
<td>65.1</td>
</tr>
<tr>
<td>RIDN-HJB</td>
<td>64.0(1.1)</td>
<td>71.5(0.7)</td>
</tr>
</tbody>
</table>

RIDN-HJB is capable of achieving the best rank-1 recognition with lowest variations.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>VR@FAR = 0.01</th>
<th>VR@FAR = 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-V2</td>
<td>68.37</td>
<td>80.46</td>
</tr>
<tr>
<td>S-V3</td>
<td>74.41</td>
<td>80.56</td>
</tr>
<tr>
<td>RIDN-ES [9]</td>
<td>86.6(0.7)</td>
<td>92.0(0.8)</td>
</tr>
</tbody>
</table>

RIDN is capable of achieving the best results on face verification, improving the true match rate on COX at a false match rate of 1% by around 12%.

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
