Accurate Periocular Recognition Under Less Constrained Environment Using Semantics-Assisted Convolutional Neural Network

Zijing Zhao, *Student Member, IEEE,* and Ajay Kumar, *Senior Member, IEEE*

***Abstract*— Accurate biometric identification under real environments is one of the most critical and challenging tasks to meet growing demand for higher security. This paper proposes a new framework to efficiently and accurately match periocular images that are automatically acquired under less-constrained environments. Our framework, referred to as semantics-assisted convolutional neural networks (SCNNs) in this paper, incorporates explicit semantic information to automatically recover comprehensive periocular features. This strategy enables superior matching accuracy with the usage of relatively smaller number of training samples, which is often an issue with several biometrics. Our reproducible experimental results on four different publicly available databases suggest that the SCNN-based periocular recognition approach can achieve outperforming results, both in achievable accuracy and matching time, for less-constrained periocular matching. Additional experimental results presented in this paper also indicate that the effectiveness of proposed SCNN architecture is not only limited to periocular recognition but it can also be useful for generalized image classification. Without increasing the volume of training data, the SCNN is able to automatically extract more discriminative features from the input data than a single CNN, therefore can consistently improve the recognition performance. The experimental results presented in this paper validate such an approach to enable faster and more accurate periocular recognition under less constrained environments.**

***Index Terms*— Periocular recognition, deep learning, convolution neural network, training data augmentation.**

I. Introduction

PERIOCULAR recognition is an emerging biometric modality that has attracted noticeable interest in recent years and a lot of research effort have been devoted to advance accuracy from the automated algorithms. The periocular region usually refers to the region around the eye, although there is no strict definition or standard from the professional bodies like ISO/IEC or NIST [41]. Periocular recognition is believed to be useful when accurate iris recognition cannot be ensured, such as under visible illumination [8], unconstrained environment [9] or when the whole face is not available, as illustrated from some real-life samples in Figure 1. It has also been shown that the periocular region is more resistant to expression variations [10] and aging [11] as compared with the face. In addition to serving as an independent biometric modality, periocular information can also be simultaneously combined with iris [2], [13] and/or face [15] to improve the overall recognition performance. However matching periocular images, particularly under less constrained environment, is a challenging problem as this region itself contains less information than the entire face and often accompanied by high intra-class variations along with occlusions like from glasses, hair, *etc*.

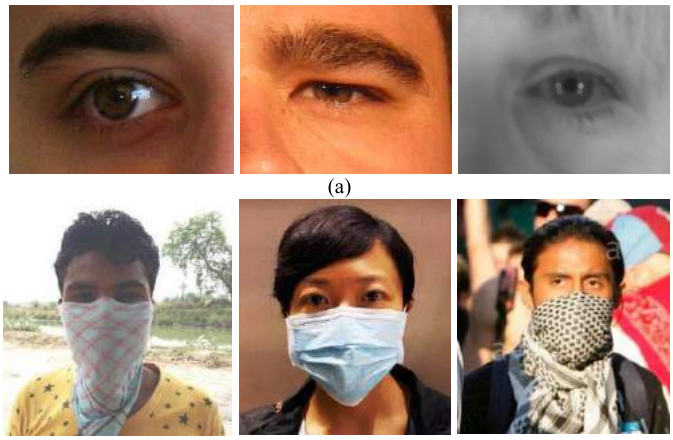


Fig. 1. Periocular recognition is useful when (a) iris texture is degraded or when the faces are covered for (b) protection from environment, (c) during sickness or (d) during demonstrations or riots [42].

In recent years, Convolutional Neural Network (CNN) has gained popularity for its strong ability to extract comprehensive features from the input data, especially for visual patterns. It has demonstrated its robustness to the real-life intra-class spatial variations. The CNN has many successful applications like hand-written character recognition [6], object detection [16], large-scale image classification [17] and face recognition [18], [19], where CNN has significantly outperformed traditional methods using handcrafted features or other learning based approaches. Therefore we have been motivated to use CNN to achieve better performance for the challenging periocular recognition problem.

Manuscript received May 30, 2016; revised October 15, 2016; accepted October 28, 2016. Date of publication December 6, 2016; date of current version February 22, 2017. This work was supported by General Research Fund from the Hong Kong Research Grant Council no. 15206814 (PolyU 152068/14E). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Venu Govindaraju.

The authors are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: jason.zhao@connect.polyu.hk; ajaykr@ieee.org).

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

Digital Object Identifier 10.1109/TIFS.2016.2636093

A. Our Work and Contributions

Automated periocular recognition under less constrained environment has shown promising performance and underlined the need for further research. Several databases, acquired under visible and near-infrared illuminations, have been introduced in the public domain [30], [34]–[36] and it can be observed that researchers require/use training samples from respective databases, primarily to select or learn best set of parameters. The performance achieved on these less-constrained databases is encouraging but requires further work. This paper attempts to address these two limitations for the automated periocular recognition.

In addition to successfully investigating the strengths of CNN for the less-constrained periocular recognition, this paper introduces the Semantics-Assisted CNN (SCNN) architecture to fully exploit the discriminative information within limited number of training samples. The key contributions of our work can be summarized as in the following.

Our approach for periocular recognition using SCNN does not require training samples from target datasets, while achieving outperforming results, which is a key advantage over state-of-the-art approaches [2] and [10]. In our experiments, the SCNN is trained with one database and tested on totally independent/separate databases. The testing and training sets have mutually exclusive subjects and highly different image quality as well as imaging conditions and/or equipment's. The SCNN architecture can also enable recovery of more comprehensive periocular features from the limited training samples. Another key advantage of proposed method in this paper is its computational simplicity, *i.e.*, our trained model requires much less computational time for feature extraction and matching compared with other methods. Unlike earlier works, the trained models and executable files of our work are made publicly available [40] so that other researchers can easily reproduce our results or evaluate on new databases. Finally, the use of SCNN is not only limited to the periocular recognition but can also be useful for general image classification task. By attaching branch CNN(s) that are trained with semantic supervision from the training data, the SCNN architecture can be easily used to extend and improve existing CNN based approaches while limiting the general requirement of increase in training data for such performance improvement. The SCNN enables the deep neural network to fully learn the training data in conjunction with the semantical correlation and therefore can benefit the final classification task, especially when the size of training data is limited to build a very deep network. The structure of SCNN is easy to implement, and semantic annotation of the training samples is often included with the release of many public databases.

B. Related Work

In 2009, Park *et al.* [8] have investigated the feasibility of using periocular region for human recognition under various conditions. Bharadwaj *et al.* [22] also support the usefulness of periocular recognition when iris recognition fails. There is also research work focusing on cross-spectrum periocular matching [5], where techniques of neural network have been used. State-of-the art work for periocular recognition includes [2], where good performance is obtained by fusing periocular and iris features/scores together. However, DSIFT feature extraction and the K-means clustering used by this work for the periocular region are highly time consuming. Another state-of-the-art approach by Smereka *et al.* [10] proposes the Periocular Probabilistic Deformation Model (PPDM), which is a variant of their previous work, and promising performance has been reported. The PPDM uses a probabilistic inference model to evaluate the matching scores from correlation filters on patches of the image pair. However, this patch-based matching scheme is sensitive to scale variance among samples, which often exists in the challenging forensic and security scenarios. More importantly, approaches in both [2] and [10] employ some samples in the target dataset for training or selection of parameters, while our objective has been to develop a more effective approach, that does not require any samples from target datasets for training, that can deliver outperforming results and is computationally simpler.

As for the development of CNN, Lecun's early work [6] for handwritten character recognition is one of the most typical applications of CNN in computer vision. Gradient based learning was used in that work so that CNN can learn from the training data effectively. In recent years CNN becomes very popular due to its powerful feature extraction ability for visual pattern and robustness for challenging scenarios (typically for large intra-class variance), and CNN based methods hold state-of-the-art performance for many computer vision tasks, such as image classification [17], object detection [16], *etc*. In 2014, Sun *et al*. [18] and Taigman *et al*. [19] have presented successful application of CNN on face recognition, which showed superior results even compared with human performance. Above two approaches have shown great potential of using CNN on biometrics, which is the primary motivation for us to develop the proposed CNN based method for the highly challenging periocular recognition problem. However, most of the existing CNN based methods require huge amount of data for training, which is the major bottleneck for its quick use for many other computer vision tasks. This has motivated us to explore alternate strategies to considerably compensate lack of large dataset often required for the training.

II. Methodology

As discussed earlier, we were motivated to incorporate CNN for the challenging periocular recognition problem due to its known ability to extract comprehensive feature from image. In this section we will first introduce the theoretical background of CNN and the practical architecture of our SCNN model in Section II.A, followed by detailing the application for the periocular recognition problem in Section II.B and II.C.

A. Semantics-Assisted Convolution Neural Network (SCNN)

1) Basic Introduction to CNN: CNN is a biologically-inspired variants of multilayer perceptron (MLP) and well-known as one of typical deep learning architectures. CNN has shown strong ability to learn effective feature representation from input data especially for image/video

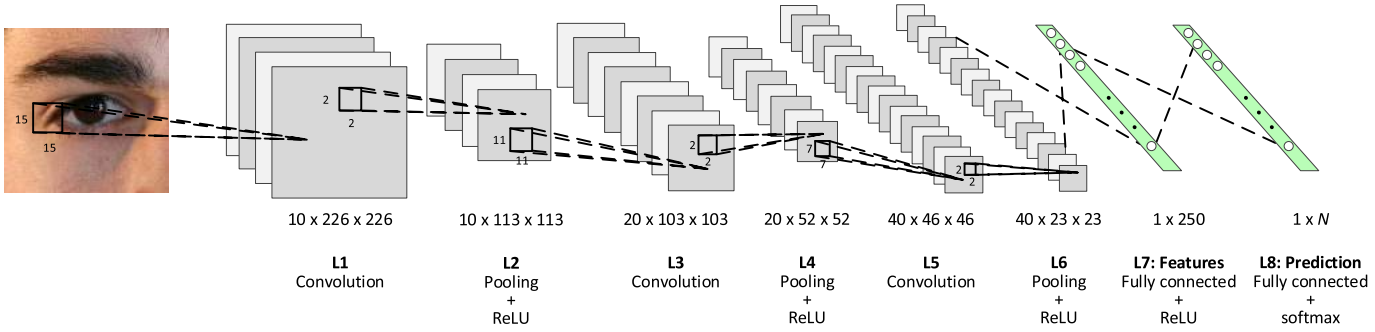


Fig. 2. Structure of the employed deep convolutional neuron network.

understanding tasks, such as handwritten character recognition [6], large-scale image classification [17], face recognition [18], [19], etc. In the following, we will briefly introduce the basic knowledge of a typical CNN architecture that is used in our and many other work.

CNN is usually composed of convolution layers, pooling layers and fully connected (FC) layers. At the output of each layer, there is often a nonlinear activate function, such as sigmoid, ReLU [1], etc. In our work, we adopt the basic CNN structure similar to AlexNet [7] and is shown in Figure 2 (say the periocular recognition problem as an example). The input image is passed through several convolutional units and then a few fully connected layers. The output of the last FC layer with N(number of classes) nodes would represent probabilistic prediction to the class labels.

Each of the convolution units is composed of three components - a convolution layer, a max-pooling layer and a ReLU (Rectified Linear Unit) activation function, as shown in Figure 2. For the convolutional layer, each channel of its output is computed as:

$$\mathbf{y}^{(i)} = \sum_j (\mathbf{b}^{(ij)} + \mathbf{k}^{(ij)} * \mathbf{x}^j) \tag{1}$$

where $\mathbf{y}^{(i)}$ is the i-th channel of the output map, $\mathbf{x}^{(j)}$ is the j-th channel of the input map, $\mathbf{b}^{(ij)}$ is called the bias term, $\mathbf{k}^{(ij)}$ is the convolution kernel between $\mathbf{y}^{(i)}$ and $\mathbf{x}^{(j)}$, and $*$ denotes the 2D convolution operation. $\mathbf{b}^{(ij)}$ and $\mathbf{k}^{(ij)}$ will be learned by back-propagation so that the convolution kernels are trained to extract most useful features that are discriminative among different subjects.

The pooling layer extracts one maximum or average value from each patch of the input channel. In our application, we use max-pooling with non-overlapping patches. As a result, the input maps, after convolution, are down-sampled with a scale determined by the pooling kernel. The pooling operation aggregates low-level features from the input to high-level representation and thus could achieve spatial invariance among different samples.

At the output of each pooling layer and the first FC layer (e.g., L7 in Figure 2), we choose the ReLU (Rectified Linear Unit) [1] as the activation function:

$$y_i' = \max(y_i, 0) \tag{2}$$

The ReLU activation ensures the nonlinearity of the feature extraction process and is more efficient for training, compared with the traditional activation functions like sigmoid or tanh employed in other approaches [14].

The FC layers process the input as in conventional neural networks:

$$y_i = b_i + \sum_j x_j \cdot w_{ij} \tag{3}$$

where x_j is the j-th element of the vectorized input map to the current layer, y_i is the i-th element of the output map, which is also a vector. b_i and w_{ij} are elements of the bias and weights to be learned through training. The last FC layer, as usually configured in classification problem, is not followed by ReLU but a *softmax* function:

$$y_i'' = \frac{e^{y_i}}{\sum_j e^{y_j}} \tag{4}$$

The use of *softmax* function in the final output of the network results in a $1 \times N$ vector with positive elements which are summed up to one. Each element then is treated as the probabilistic prediction of the class label. The cross-entropy loss function is to be minimized, which is formulated as:

$$L(\mathbf{y}'') = -\log y_t'' \tag{5}$$

where t is the ground truth label of the training sample. The loss function is minimized via back-propagation so that the predictions of the ground truth class of the training samples will approach to unity.

2) Limitation of Contemporary CNN Based Approach: In order to achieve superior performance using CNN based methods, a common way is to add more layers to make the network deeper and more comprehensive, and/or devote more labeled training data because CNN is usually trained in a supervised manner. For instance, the famous CNN architecture GoogLeNet [17] has 22 layers and later comes the Microsoft's deep network with 152 layers [21]. Apparently, common researchers or companies could hardly afford to train such deep networks due to the lack of enough computational power. Also, as the network goes deeper, the need for training data grows accordingly, while in many research areas, it is difficult to acquire enough labeled training samples like ImageNet [23]. Table I provides examples of several typical deep learning

TABLE I

EXAMPLES OF SEVERAL DEEP LEARNING BASED APPROACHES AND THEIR REQUIRED NUMBER OF TRAINING IMAGES

Approach	Task	Size of Training Data	
		No. of Classes	No. of Samples
CVPR [17]	Image classification	1,000	1,281,167
ICML [24]	Handwritten digits recognition	10	60,000
T-PAMI [25]	Object detection	200	456,567
CVPR [18]	Face recognition	10,177	202,599
CVPR [19]	Face recognition	4,030	~ 4,400,000

based approaches and their employed training data. In reference [18, Table 1], for instance, where the developed CNN is not very deep (nine layers), a total of ~200,000 face images from more than 10,000 people were used for training to achieve superior performance. However for other popular biometrics modalities like iris or periocular, in the best of our knowledge, there is currently no single public database with that many images.

Therefore, we are motivated to improve the performance of existing CNN based architecture in another way - to enhance CNN with supervision from explicit semantic information. When human recognizes objects, for example while recognizing a face image, one would analyze not only the overall visual pattern but also the semantic information, such as gender, ethnicity, age, *etc*., to judge whether the face image belongs to a certain known person. Therefore, it is reasonable to believe that semantic information is helpful for the visual identification task. For a CNN that is trained with the identity label only, it is possible that the network is already capable of acquiring semantic information. For instance, for the well-known deep learning model for face recognition, *DeepID2+*, researchers discovered that although the network was trained using subject identities, certain neurons turn out to exhibit selectiveness to attributes like gender, ethnicity, age, *etc.* These semantic attributes contribute to discriminating identities [43]. However, such useful semantic information is expected to be *implicitly* learned by the CNN. It is not easy to answer the following questions:

1) How many types of semantic information can be acquired? Since the discriminative capacity of a certain CNN is limited, we cannot guarantee that all the semantic information we prefer to have has already been included.
2) To what extent the semantic information can be analyzed by the trained CNN? Does it really help in the final identification task, or could it be further improved?

Above problems arise due to the nature of training popularly employed for the CNN, *i.e*., the loss function is usually only related to the class labels, therefore it is hard to reveal how the semantic information can be *implicitly* acquired. In order to address this issue, we propose to empower the CNN with the ability to analyze semantic information *explicitly*. The idea is very simple and illustrated in Figure 3.

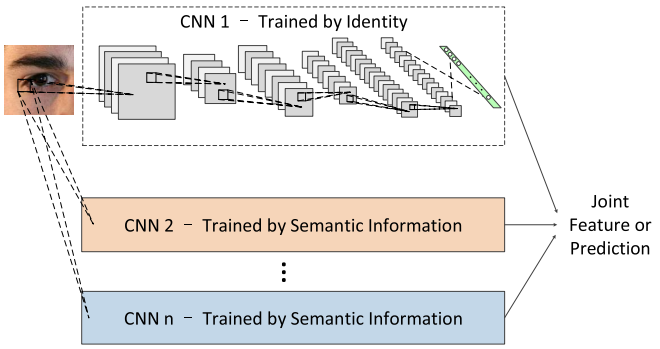


Fig. 3. Structure of the proposed Semantics-Assisted CNN (SCNN). While first branch is trained by the label of the intended tasks, other branch CNNs are trained using different semantic information, then the branches are joint in the end to get a comprehensive feature representation or perform score fusion.

3) Semantics-Assisted CNN: As illustrated in Figure 3, we simply add a branch, which is also a CNN, to the existing CNN. The attached CNN is not trained using the identity of the training data but the semantic groups. For example, we could train CNN2 using the gender information of the training sample, *i.e*., let the CNN2 be able to estimate the gender instead of identity, and train CNN3 using the ethnicity information. After the CNNs are trained, we can combine the output of each CNN in the way of feature fusion. We refer to such extended structure of the CNN as *Semantics-Assisted CNN* (*SCNN* for short). Despite the simplicity of this idea, it can inherently improve the original CNN by adding more discriminative power to it, which has been shown from the experiments described in Section 3. Theoretically, the SCNN has the following benefits:

- Instead of letting the semantic information be learned from the identities by the CNN in an unpredictable and uncontrollable way, SCNN allows us to *explicitly* recover the preferred semantic information that can be helpful for the identification task. As a result, the feature representation from the SCNN is accompanied by more reliable semantic information that is closer to mechanism in human visual system.
- The training scheme for SCNN can reuse the same set of training data but just labeled in another way than the simple identities. Since the labeling scheme is variable, the branches of SCNN learn the training data from different points of view, which is equivalent to increasing the data volume without really adding the number of training samples. This can relax the constraints on the requirements of enormous training data for deep neural networks to some extent, *i.e*., instead of pursuing for superior performance from a single CNN, we enhance the joint performance of branches of CNNs with fewer amounts of training data.
- The SCNN architecture and training scheme is naturally compatible for most of the existing CNN based approaches. What we need is just to train some independent CNNs with semantic grouping labels and judiciously combine the features from multiple CNNs to benefit from such training, as the semantic annotations of training

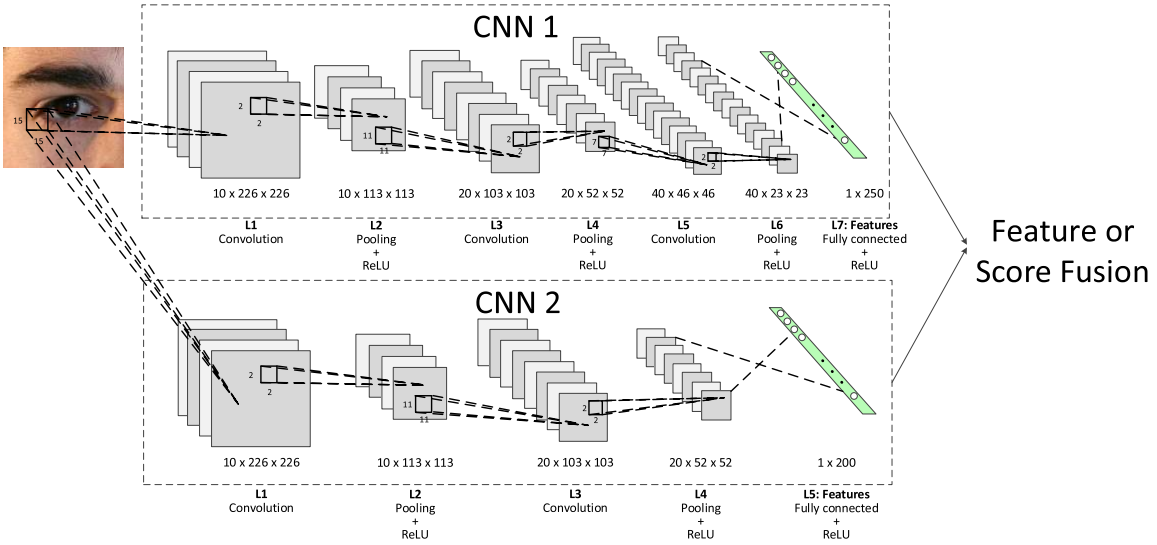


Fig. 4. Structure of the employed SCNN for the periocular recognition.

samples are also available for many public databases. In addition, the architecture of SCNN is highly friendly for parallel computing platforms.

B. Application for Periocular Recognition

As discussed earlier, CNN has been successfully used for the face recognition in several state-of-the-art approaches [18], [19]. Considering that the periocular region is actually a part of face and also presents some structural information (eyebrow, eyelids, eyeball, *etc*.), it is reasonable to expect that CNN can be effective for the periocular recognition problem. However, as compared with such related work, we are constrained by lack of large-scale periocular databases that are usually required to sufficiently train a deep neural network. Therefore we developed and investigated SCNN for the periocular recognition problem.

1) Network Structure and Supervision Information: The detailed SCNN structure used for the periocular recognition is shown in Figure 4. In order to examine the impact of adding branch to an existing CNN, we simply designed one branch that is trained with semantic information, denoted as CNN2 in Figure 4. While CNN1 is like the ones commonly trained with the subject identities from the training samples, CNN2 is designated to be trained with the side (left or right) and the gender information. More specifically, we labeled the training data as follows, also shown in Figure 5:

$$\begin{cases} 0-\textit{Left and Male}, \\ 1-\textit{Right and Male}, \\ 2-\textit{Left and Female}, \\ 3-\textit{Right and Female}. \end{cases}$$

The reason for using left/right and gender information is that humans also tend to incorporate such judgment by visually inspecting the presented periocular images, although such accuracy may not reach cent percent level. Therefore there is some scientific basis to believe that CNN can learn to distinguish above semantic information from the periocular patterns and assist in the identification task. Another reason for using gender information is that the genders of subjects are often included in the metadata of many publicly available datasets, such as UBIpr [36]. Therefore we can directly use those labels to train CNN2. Other possible and useful semantic information include iris color (light/dark), ethnicity, shape of eyebrow, *etc*.

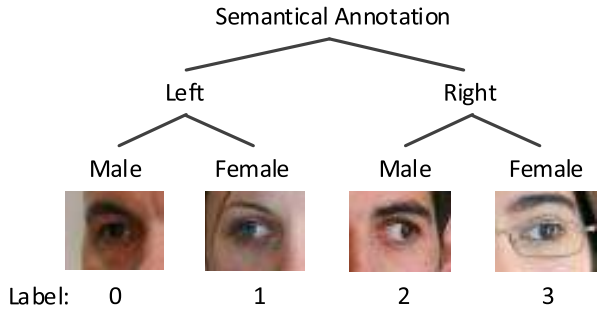


Fig. 5. Semantical labeling used in our implantation to train CNN2.

Using such additional semantic information to supervise the network makes the overall architecture and learning process of SCNN similar to multi-label learning [44] to some extent. However, the principal difference is that, the introduction of semantic labeling in our model aims to assist/supplement the prediction of subject identity labels, *i.e.*, they are inequally important, while in traditional multi-label learning, the multiple labels are usually in equal positions. In addition, the learning processes of identities and other semantic information are separately undertaken to maximally ensure the explicitness of semantic learning and compatibility to other CNN based model, while in general multi-label learning, features are usually jointly learned for predicting different lables. Nevertheless, in spite of the diffrentiation between the identity labels and other supportive labels, the semantic learning process

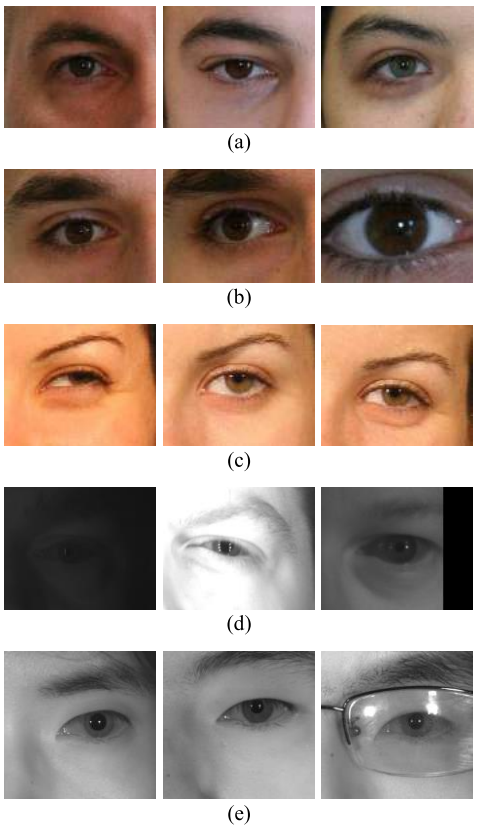


Fig. 6. Sample images from the databases we used in the experiments. Scale variance and misalignment are common in the testing environment. (a) UBIpr (training). (b) UBIRIS.V2 (testing). (c) FRGC (testing). (d) FOCS (training and testing). (e) CASIA.v4-distance (testing).

(*e.g.*, CNN2 itself) can also be conducted in the manner of multi-label learning alternatively.

2) Training Protocol and Data Augmentation: Among the original training samples, the last sample of each subject is selected to form the validation set, which is tested in every certain amount of iterations to observe whether the training process is converging in a right direction or not.

Furthermore, it is observed that the periocular images from the training set are well aligned and scaled to a similar level, while the samples from independent test datasets and real applications may have misalignments and scale variations. Such inconsistency can also be observed from the image samples in Figure 6.

If the deep network is trained with the well aligned and scaled images, it may not be effectively generalized to other datasets or data acquired by real applications. In order to address such problems, we firstly augmented the training data with a different scale to simulate scale inconsistency in the test environment. Then we applied random cropping during the training process to ensure that the network can accommodate spatial variations among the periocular images. The scale augmentation and random cropping process is also illustrated in Figure 7. As illustrated in this, each original of the image in training set is automatically cropped from its center with a size of $0.6w \times 0.6h$, where w and h are its original width and height respectively. The original images and its cropped patch are resized to300×240, then padded with symmetric edges

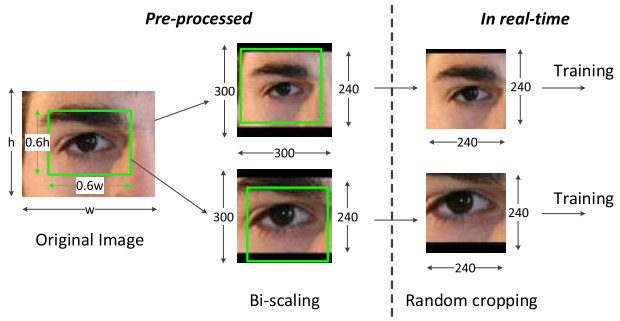


Fig. 7. Illustration of scale augmentation and random cropping. Each original image is augmented to two samples with different scales, and each augmented sample would be cropped by a smaller window that is randomly placed before entering the network for each epoch of the training process.

filled with zeroes to a size of 300×300. So far one original periocular image could generate two training samples. As a result, we have 6,270 samples for training and 448 samples for validation while training for each side of the periocular images. Furthermore, during the training process, each training sample would be cropped by a 240×240 window randomly placed within the image region before entering the first layer of the network. Such randomized cropping process from one training sample could produce abundant samples that have randomized misalignments with others. In this way, the network can be enforced to learn to extract features that are robust to the misalignments.

3) Visualization of Trained SCNN: Once the networks have been trained, CNN1 is expected to lock-into features that are directly relevant to the subject identities, while CNN2 is expected to analyze the features that are more related to side and the gender difference. In order to observe the difference among features extracted by the two CNNs, we have visualized the filter kernels from the first two convolutional layers of trained CNN1 and CNN2 in Figure 8.

We can visually observe from Figure 8 that: 1) Overall both CNNs were not trained sufficiently. Compared with convolutional kernels trained with large amount of samples (*e.g.*, those in [7]), a number of kernels here remain flat or noisy, for which it is less likely to extract useful information. Insufficiently trained network parameters usually results in certain levels of over-fitting. 2) Despite the over-fitting concern, the convolutional filter kernels of CNN1 and CNN2 are quite different. Critical kernels in CNN2 are sharper and present more visual salience, therefore might be more sensitive to small texture, edges or corners than the filters in CNN1. This indicates CNN2 can provide complementary information that CNN1 was not able to learn due to lack of sufficient training data. Although the features extracted by CNN2 are not directly related to the subject identities, it is reasonable to expect that those visual features could assist CNN1 to form a more comprehensive visual representation of the periocular image, therefore help to distinguish different subjects finally.

C. Feature Vector and Verification Score Generation

The CNNs we use are trained in a classification protocol, *i.e.*, the category or identity of the input data is known and fixed. Therefore this network can be directly used in some

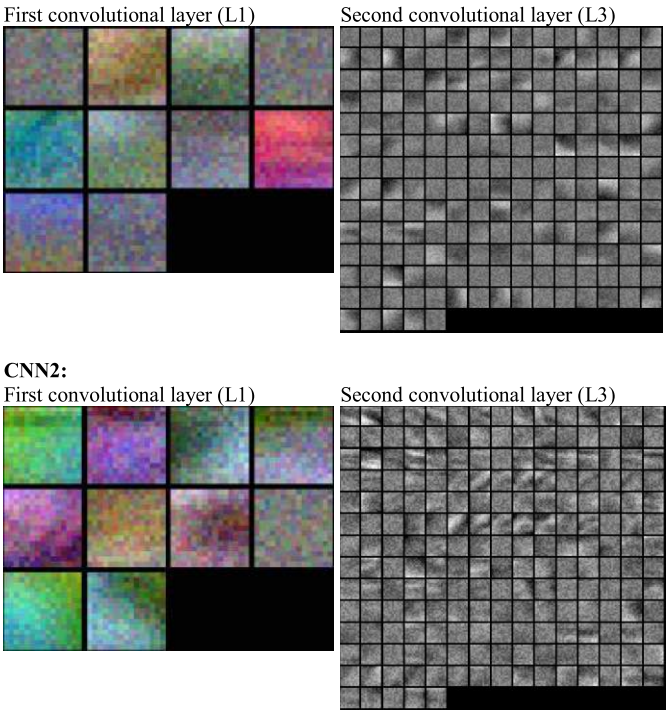


Fig. 8. Visualization of the filter kernels from the first two convolutional layers of trained CNN1 and CNN2 respectively.

classification or identification tasks. However, in biometrics, one-to-one matching for probably unseen subjects is the key problem and needs to be evaluated. Therefore, we need to generalize the trained model to separated subjects that are not included in the training set, and formulate one-to-one matching scheme.

Similar to [18], we use the output of second last layer (L7 in CNN1 and L5 in CNN2) as the feature representation of the input data. While the last layer represents the class prediction during the training process, the second last layer should contain the most relevant and aggregated information that can contribute to distinguishing the classes or identities. Therefore, it is reasonable to use the output of the second last layer as the feature representation and generalize the model to unseen subjects. Once we get the layer output vectors, we first normalize them by l^2 norm, then apply PCA to reduce the dimensionality of the vector. For the SCNN architecture, we simply concatenate the two independently normalized output vectors to form a longer vector before PCA. In our experiments, the dimension of output vectors after PCA is set to 80, for both the single CNN and SCNN cases. Then the joint Bayesian scheme [33] is utilized to predict the similarity between a pair of feature vectors. The joint Bayesian is primarily designed for face verification, in which a face (equivalent to the periocular feature vector here) is represented by:

$$\mathbf{f} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{6}$$

where $\mathbf{f}$ is the observation, in this paper the feature vector after PCA, $\boldsymbol{\mu}$ is the identity of the subject, $\boldsymbol{\varepsilon}$ is the intra-class variation. $\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}$ are assumed to be two independent Gaussian variables following $N(\mathbf{0}, \mathbf{S}_\mu)$ and $N(\mathbf{0}, \mathbf{S}_\varepsilon)$ respectively, then the covariance of two observation is:

$$cov(\mathbf{f}_1, \mathbf{f}_2) = cov(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) + cov(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2) \tag{7}$$

The joint distribution of a pair of observations $\{\mathbf{f}_1,\mathbf{f}_2\}$ is considered. Let H_I denote the intra-person hypothesis indicating that two observations are from the same person, and H_E the extra-person hypothesis. Under H_I, since $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the same, $\boldsymbol{\varepsilon}_1$and $\boldsymbol{\varepsilon}_2$ are independent, the covariance matrix of the distribution $P(\mathbf{f}_1,\mathbf{f}_2|H_I)$ is:

$$\boldsymbol{\Sigma}_I = \begin{bmatrix} \mathbf{S}_\mu+\mathbf{S}_\varepsilon & \mathbf{S}_\mu \\ \mathbf{S}_\mu & \mathbf{S}_\mu+\mathbf{S}_\varepsilon \end{bmatrix} \tag{8}$$

On the other hand, under H_E, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are also independent, therefore the covariance matrix has become:

$$\boldsymbol{\Sigma}_I = \begin{bmatrix} \mathbf{S}_\mu+\mathbf{S}_\varepsilon & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_\mu+\mathbf{S}_\varepsilon \end{bmatrix} \tag{9}$$

With above conditional joint probabilities, the log likelihood ratio which tells the difference between intra- and extra-person probabilities can be obtained in a closed form:

$$r(\mathbf{f}_1, \mathbf{f}_2) = \frac{P(\mathbf{f}_1,\mathbf{f}_2|H_I)}{P(\mathbf{f}_1,\mathbf{f}_2|H_E)} = \mathbf{f}_1^{\mathrm{T}}\mathrm{A}\mathbf{f}_1 + \mathbf{f}_2^{\mathrm{T}}\mathrm{A}\mathbf{f}_2 - 2\mathbf{f}_1^{\mathrm{T}}\mathbf{G}\mathbf{f}_2 \tag{10}$$

where

$$\mathrm{A} = (\mathbf{S}_\mu + \mathbf{S}_\varepsilon) - (\mathbf{F} + \mathbf{G}) \tag{11}$$

$$\begin{bmatrix} \mathbf{F}+\mathbf{G} & \mathbf{G} \\ \mathbf{G} & \mathbf{F}+\mathbf{G} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_\mu+\mathbf{S}_\varepsilon & \mathbf{S}_\mu \\ \mathbf{S}_\mu & \mathbf{S}_\mu+\mathbf{S}_\varepsilon \end{bmatrix}^{-1} \tag{12}$$

The covariance matrix $\mathbf{S}_\mu$ and $\mathbf{S}_\varepsilon$ can be estimated using an EM based algorithm as detailed in [33], and the log likelihood ratio $r(\mathbf{f}_1, \mathbf{f}_2)$ is used as the similarity score in our one-to-one matching scenario.

III. Experiments and Results

In this section we provide the details on the experiments and analyze the results. The experimental details on the periocular identification are firstly provided and this is followed by details on supporting experiments for the image classification.

A. Periocular Recognition

1) Training and Testing Datasets and Protocol: We use following publicly available databases for the experiments. Two different databases were employed for training the deep neural networks and three separate databases were employed for the testing.

- UBIpr [36] - *for training*

We employed UBIpr periocular database [26] for training the SCNN for the visible spectrum. This database originally contains 5,126 images for each of left and right perioculars from 344 subjects. However, we are also employing a subset of UBIRIS.v2 database [4] for separate test experiments, which has some overlapping subjects with the UBIpr database. In order to ensure that subjects of training set and testing set are mutually exclusive, we removed these overlapping subjects from UBIpr database before we perform training on the network. As a result, we only have 3359 periocular images from

each of the two sides of 224 subjects. Such a scale is relatively small as compared with those in the training protocols in other typical deep learning work like ImageNet [27] or LFW [28]. Therefore, the application scenario is good for validating the ability of SCNN for learning comprehensive information from limited size of training data.

- UBIRIS.v2 [4]

The UBIRIS.v2 database is primarily released for evaluation of at-a-distance iris segmentation and recognition algorithms under visible illumination and challenging imaging environment. Since the eye images in this database contain surrounding regions of the eye, it is possible to perform periocular recognition on the UBIRIS.v2 database. Similar to as in [2], we use a subset of 1,000 images from this database that is released in NICE.I competition [29]. This subset contains left and right eye images together from 161 subjects that are captured from 3m to 8m, bringing serious scale inconsistency. Some images only contain the eye region without eyebrow and other surrounding texture which makes the task of periocular recognition highly challenging. Some sample images are shown in Figure 6(b).

- FRGC [30]

The dataset of Face Recognition Grand Challenge (FRGC) is released by the National Institute of Standards and Technology (NIST) and has been primarily for the evaluation of new algorithms for the automated face recognition. Similar to as in [2], we automatically extracted the periocular region from the original face images of FRGC using publicly available face and eye detector [31], [32]. A subset of 540 right eye images from 163 subjects, same as also the ones used in [2], were employed in the experiments. Some sample images are reproduced in Figure 6(c).

- FOCS [34] - *for training and testing*

The Face and Ocular Challenge Series (FOCS) dataset is also released by NIST and contains face, ocular images and videos. We employed the "OcularStillChallenge1" section, which consists of 4,792 left and 4,789 right periocular images from 136 subjects that are cropped from face video clips acquired under near-infrared (NIR) spectrum. The periocular samples from this dataset, as shown in Figure 6, suffer from serious illumination inconsistency and misalignments, therefore this dataset is considered as highly challenging. We used 3,262 left and 3,259 right periocular images of the first 80 subjects to train the CNNs and used the remaining images from 56 subjects for testing. Again, such a scale of training samples and subjects is small compared with other typical deep learning tasks.

- CASIA.v4-distance [35]

CASIA.v4 is the first publicly available long-range iris and face database acquired under NIR illumination, which is released by the Center for Biometrics and Security Research (CBSR) from the Chinese Academy of Sciences (CASIA). The full database contains 2,567 images from 142 subjects in single session. The standoff distance of the subjects to the camera is from 3 meters away. Similar to FRGC, we used publicly available eye detector [31], [32] to automatically segment left periocular images which are used in our experiments. The first eight samples of each subject, excluding a few badly segmented images, were used for the periocular matching experiment.

Above datasets were selected for evaluation because of the availability of periocular images acquired under less constrained environments that are close to real world scenarios. The selected subsets from FRGC and UBIRIS.v2 contain multi-session data and exhibit obvious scale/illumunation variation. Samples in FOCS database suffer from significant illumination degradation and misalignment. Images from CASIA.v4-distance are more consistent than the other three databases, but were acquired at a distance and some contain artifacts like glasses and/or hair, therefore also represent less constrained scenarios. In addition, networks for visible and NIR spectrums were trained separately due to the significant difference between the image properties.

It is important to clarify that during our (reproducible [40]) experiments, the SCNN is tested in totally *cross-database* manner, *i.e.*, not only the subjects from the training and test set sets are totally separated, the databases themselves are independent from training for *three sets* of experiments. However, the methods we are going to compare with, [2] and [10], both require some samples of the target databases for the training. In order to compare with the best performance of [2] and [10] as well as to ensure the fairness in such comparison, we still divide the target datasets into training and testing sets, as summarized in Table II. For example, 96 samples of the first 19 subjects in UBIRIS.v2 were used to train the models [2] and [10], the remaining were used for test as in [2] and [10] and also for our method. Such a configuration is highly disadvantageous to our methods because the inter-database variance is always a key factor for the performance of all learning based methods. However, our method has still been able to achieve outperforming results as detailed later.

We perform periocular matching using the all-to-all protocol, *i.e.*, every image is matched to all the other images in the testing set, and all the generated matching scores are taken into calculation of the receiver operating characteristic (ROC) curve. Such a protocol is considered to be highly challenging because one bad sample may result in several poor genuine scores, which drops the overall matching performance.

2) Effectiveness of SCNN: We firstly examine the impact of the added branch that has been trained with the semantic information. We have compared the performance of a single CNN, *i.e.*, only CNN1 in Figure 3, with the performance of the extended SCNN. The results from the verification experiments are illustrated in Figure 9.

We can observe from Figure 9 that the SCNN consistently achieves better performance than that of original or single CNN. This observation suggests that the newly added CNN2 which is trained with semantic supervision has been successful in contributing to some useful information that is not reinforced in CNN1, and therefore improving the overall discriminative power of the network. In theory, we can add more branches that are trained with different semantic information (*e.g.*, iris color) to further improve the final recognition accuracy. However, the need for computational power would also increase and the trade-off may need to be made according

TABLE II

SUMMARY OF THE EMPLOYED DATABASES FOR TRAINING AND TESTING

Spectrum	Visible			Near Infrared (NIR)		
Division	Training set	Testing set		Training set	Testing set	
Dataset	UBIpr	UBIRIS.v2	FRGC	FOCS	FOCS	CASIA.v4-distance
Standoff distance	4 – 8m	3 - 8m	N/A	N/A	N/A	$\geqslant$3m
No. of subjects*	224	171(19/152)	163 (13/150)	80	56	141 (10/131)
No. of images*	left: 3,359 right: 3,359	1,000 (96/904)	540 (40/500)	left: 3,262 right: 3,259	1,530	1,077 (79/998)

* In the bracket (a/b) means a subjects or images were used for training for methods [2] and [10] (not for our method), remaining b subjects or images were used for testing.

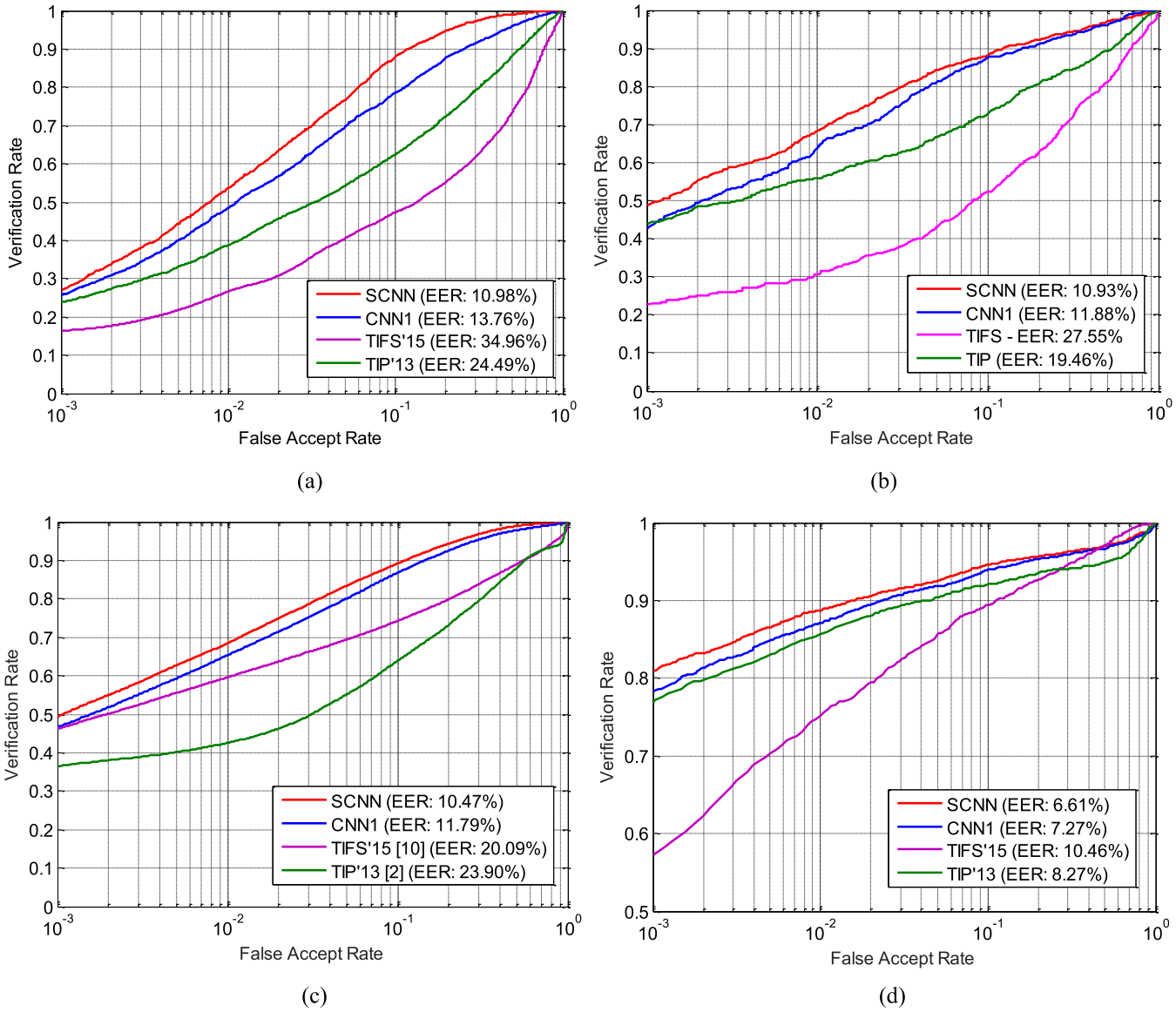


Fig. 9. ROC curves of the periocular verification using SCNN and comparison with single CNN and other state-of-the-art methods for different databases. (a) UBIRIS.v2. (b) FRGC. (c) FOCS. (d) CASIA.v4-distance.

to the applications. In our example, since CNN2 shown in Figure 4 has a relatively simplified structure, the additional training cost is minor.

3) Comparison With Earlier Work on Periocular Recognition: We also compared the performance of our approach with state-of-the-art approaches [2], [10] on the periocular recognition problem. While [2] is our previous work, we have carefully implemented the methods in [10] with the help of the original authors. The test protocols were kept exactly the same for different approaches during the experimental process

and therefore the comparisons of ROC/CMC curves are fair. However several factors can be firstly clarified here to ensure clarity in understanding the experimental comparisons.

1) For UBIRIS.v2, we use the 1,000 image set that was employed for the NICE.I competition. This subset is the same as was used in [2] but different from the one in [10]. In [10], test images were gathered from the full dataset, but only those acquired from 6-8 meters were used, while the 1,000 image set in [2] included samples acquired from 3-8 meters. Due to the relatively consistent imaging distance, the subset used in [10] involves much less scale variance than those in [2] and also in this paper. As a result, the performance from our experiment using exact method in [10] is not reproduced as good as what appears to be in [10] and this is reasonable due to the difference in selection of images as explained above.

2) For FRGC, we also used the same subset as in [2] but different from the one used in [10]. As described before, the subset we used contains 540 periocular images which were *automatically* segmented from the original face images and therefore may suffer from some misalignment. Moreover, images in this subset were acquired from *various sessions* with certain time lapse and different imaging environments, which increases the difficulty for accurate recognition. However, the subset used in [10] only consists of images captured in consistent illumination and background in *single session*, and the periocular regions were *manually* segmented. Therefore, it is also a reasonable explanation for the drop in performance in our reproduced results, over the ones shown in [10] using manual segmentation.

3) For FOCS, we used fixed division of training and testing sets as shown in Table II, while the original setup in [10] used 5-fold cross validation for the entire dataset. Although the subsets used in our experiment and their original experiment are not exactly the same, the quality of images is observed to be quite similar. Therefore our reproduced result is very close to those appearing in [10].

The verification results (ROC) for above comparisons are also shown in Figure 9, while the identification results (CMC) are shown in Figure 10. It can be observed from the experimental results in these two figures that the proposed approach using SCNN consistently outperforms the two state-of-the-art approaches.

In order to ascertain statistical significance of the improvements, we have conducted the significance test for the ROC curves using the method described in [12], which judges from the area under the curve (AUC). Table III shows the significance level (p-value) of the difference of the SCNN based method over the comparative methods [2] and [10]. The results indicate that, by the commonly used confidence level of 95%, our approach significantly outperforms these two methods (p-value < 0.05) on all the employed datasets.

It may be noted that [10] performed poorly on the UBIRIS.v2 set because it adopts the patch based matching scheme while, as explained above, the 1,000-image set of UBIRIS.v2 used in our experiment suffers from serious scale variations among the samples, which results in significant loss of patch correspondence. The approach from [2] which uses DSIFT features is more robust to scale variance, however the extraction of DSIFT feature is especially time consuming. In contrast, our approach not only performs better than both of the baseline approaches on different databases, but is also computationally simpler for the deployment using the trained network. Table IV presents the summary of the average time required for the feature extraction for the considered state-of-art approaches. These tests were performed using the Matlab wrapper and C++ implementation running on a computer with Linux OS, 16 GB RAM, 3.4 GHz Intel®Core™i7-4770 CPU (4 cores) and NVIDIA®GeForce GTX 670 GPU. It can be observed that the proposed approach is much faster due to the straightforward architecture and the use of GPU could further reduce the computational time.

TABLE III

RESULTS OF SIGNIFICANCE TEST FOR COMPARISON OF ROCS USING METHOD [12]. p-VALUE INDICATES THE PROBABILITY OF THE NULL HYPOTHESIS THAT TWO METHODS HAVE NO DIFFERENCE STATISTICALLY

Comparison	p-value			
	UBIRIS.v2	FRGC	CASIA.v4 -distance	FOCS
SCNN & TIP'13 [2]	< 1e-4	< 1e-4	< 1e-4	< 1e-4
SCNN & TIFS'15 [10]	< 1e-4	< 1e-4	< 1e-4	< 1e-4

* The computed z-statistics are too large that the corresponding p-values exceed double precision, therefore expressed as < 1e-4.

TABLE IV

COMPARISON OF TIME REQUIRED TO MATCH TWO PERIOCULAR IMAGES BY DIFFERENT APPROACHES, FROM MATLAB IMPLEMENTATION RUNNING ON A COMPUTER WITH LINUX OS, 16 GB RAM, 3.4 GHz INTEL I7-4770 CPU (4 CORES) AND NVIDIA GEFORCE GTX 670 GPU

Approach	Major Time Consuming Operations	Matching Time (s)	
		GPU	CPU
proposed	convolution, matrix multiplication	**0.013**	**0.183**
TIP'13 [2]	DSIFT feature extraction, K-means clustering	/	15.478
TIFS'15 [10]	Gabor feature extraction, correlation filter matching	/	1.441

B. Image Classification

In order to examine that the proposed SCNN architecture is not only effective for the periocular recognition but can also be useful for more general problems, we performed experiment for image classification on the CIFAR-10 dataset [37].

The CIFAR-10 dataset contains 60,000 32×32 color images from 10 classes. Among these images, 50,000 images are for training and 10,000 are for testing. Figure 11 shows some randomly selected samples from each class. As we can see from Figure 11, although the number of classes is not large, the intra-class variation is significant and the resolution is also smaller, which brings certain challenge for classifying those images. The CIFAR-10 has therefore emerged as a popular dataset for evaluating image classification algorithms along with others like ImageNet and CIFAR-100, *etc*.

Since the SCNN is developed to enhance existing CNN based approaches, we select a baseline CNN to ascertain

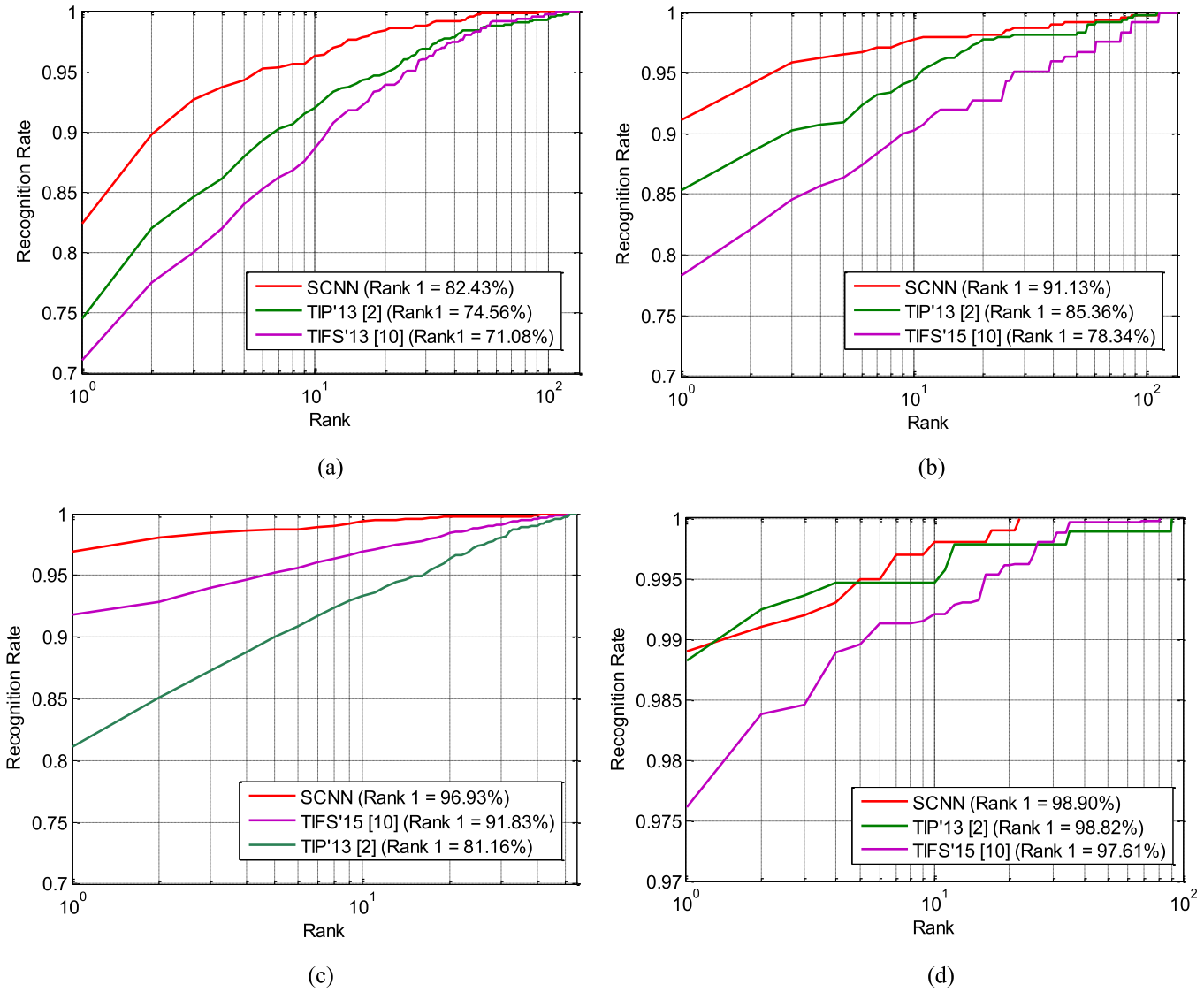


Fig. 10. CMC curves of the periocular verification using SCNN and comparison with state-of-the-art methods for different databases. (a) UBIRIS.v2. (b) FRGC. (c) FOCS. (d) CASIA.v4-distance.

the improvement. We adopt the CNN originated from Krizhevsky's cuda-convnet [38], re-implemented and introduced in the Caffe tutorial [39]. Although the selected CNN is not the state-of-the-art for CIFAR-10 in terms of performance, we chose it because this model is publicly available under Caffe, the deep learning framework employed in the paper, and it is also quick to train. For simple annotation, we refer to this network as *cuda-convnet*. By following the tutorial, we can quickly get an accuracy of about 75% on the CIFAR-10 test set. Then we trained a branch CNN to learn the semantic features of the images in CIFAR-10 in order to build the SCNN architecture. We define one possible groups of semantic information for the classes in the CIFAR-10 dataset as follows, also shown in Figure 12.

- artificial
  - rectangular, has wheel: (automobile, truck)
  - no/invisible wheel: (airplane, ship)
- Natural
  - round, short: (cat, dog, bird, frog)
  - slim, long: (deer, horse)

With above division, the entire dataset is grouped into four semantical classes. It may be noted that this is not the unique or the optimal division, but it is an easy-to-understand scheme to start with. In order to obtain a branch CNN that was trained to acquire above semantic features, we simply duplicate the structure of the base cuda-convnet but replace the last fully connected layer having 10 neurons with a new fully connected layer with four neurons, since the task now is to recognize the four semantic groups. We then just repeat, as described in Caffe tutorial, but train the new network with newly labeled data. We refer to this new CNN as cuda-convnet-s. Again, above configuration is made because of the ease to execute and one has many choices for actual applications. We then built an SCNN with the architecture as in Figure 13. As shown in this figure, we combine the branch CNN and the original one to obtain an extended structure. The components highlighted in red are retrained after the combination to aggregate the long concatenated features, and this process can be considered as a kind of finetuning. Since the number of layers to be

Fig. 11. Sample images from each class of CIFAR-10 dataset.

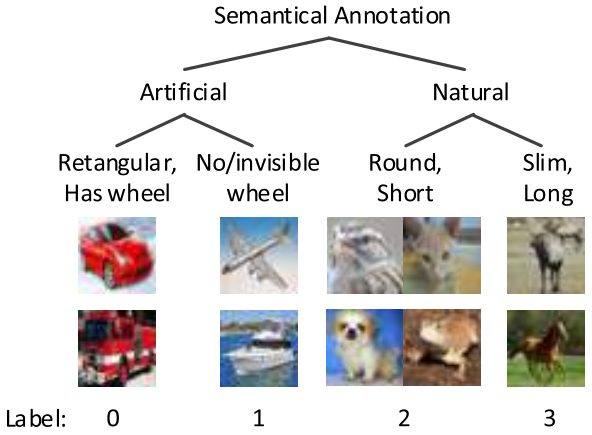


Fig. 12. The semantical group labeling used in our experiment to train cuda-convnet-s.

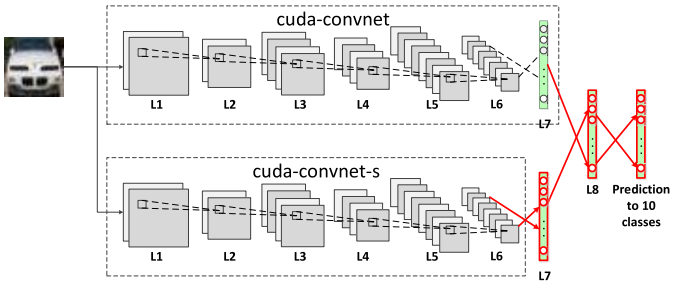


Fig. 13. The structure of SCNN used in the experiment for CIFAR-10 dataset. The *cuda-convnet* is from the original Caffe tutorial, and the *cuda-convnet*-s is newly trained by the semantic information.

retrained is small, the finetuning is very fast. Table V shows the classification results on the test set using the original cuda-convnet and the extended SCNN.

TABLE V

RESULTS OF CLASSIFICATION ON THE CIFAR-10 TESTING SET USING ORIGINAL EXISTING *cuda-convnet* AND THE PROPOSED SCNN ENHANCEMENT ON THE *cuda-convnet*

Approach	Accuracy
cuda-convnet	74.95%
cuda-convnet-SCNN	77.06%

We can observe from the results that the proposed SCNN can achieve an improvement of 2.11% over the original result. Although this may not be considered as a very large improvement, the achieved results reinforce the motivation for SCNN is to make solid and consistent enhancement on existing CNN based approaches, especially for the scenario when the training data may not be enough to feed a complex network. In the CIFAR-10 dataset, the number of images per class is actually quite large and therefore the effect of SCNN is not significant, but it still offers a noticable improvement with minor addition in the complexity. Moreover, as discussd above, the experimental setup is reproducible and made to execute in a straightforward manner. Therefore it is reasonable to expect certain space for further improvement.

IV. CONCLUSIONS

This paper has presented automated periocular recognition using CNN with outperforming results and significantly smaller complexity. In particular, we proposed a robust and more accurate framework for the periocular recognition using the semantics-assisted convolutional neural network (SCNN). By training one or more branches of CNNs with semantical information corresponding to training data, the SCNN is capable of recovering more comprehensive features from the images and therefore achieve superior performance. Our experimental results on four publicly available databases suggest that the proposed approach can achieve outperforming results while requiring much smaller computational time for the matching process. The SCNN architecture can also be generalized for other image classification tasks, which can improve the performance over the single CNN based approaches. The source and executable files of our approach are made publicly available [40] to encourage other researchers to easily reproduce our results and further advance research on accurate periocular recognition.

It may be noted that at the current stage, we decouple the identity supervision and other semantic supervision, in order to ensure high level of explicitness of semantic learning and compatibility to existing CNN based approaches. However, it is believed that a well-designed network structure may explicitly incorporate semantic information itself and facilitate efficient training in an end-to-end training manner. It will be our future work to investigate improved architecture which enables joint learning of semantic information explicitly as well as preserving the network integrity.

ACKNOWLEDGEMENT

The authors thankfully acknowledge the support and help received from the authors of reference [10] in reproducing results from their approach employed for comparisons in this paper.

REFERENCES

[1] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[2] C.-W. Tan and A. Kumar, "Towards online iris and periocular recognition under relaxed imaging constraints," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3751–3765, Oct. 2013.

[3] L. Nie, A. Kumar, and S. Zhan, "Periocular recognition using unsupervised convolutional RBM feature learning," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, Stockholm, Sweden, Aug. 2014, pp. 399–404.

[4] H. Proenca, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre, "The UBIRIS.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1529–1535, Aug. 2010.

[5] A. Sharma, S. Verma, M. Vatsa, and R. Singh, "On cross spectral periocular recognition," in *Proc. Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 5007–5011.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[8] U. Park, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum: A feasibility study," in *Proc. IEEE 3rd Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, Sep. 2009, pp. 1–6.

[9] G. Santos and H. Proenca, "Periocular biometrics: An emerging technology for unconstrained scenarios," in *Proc. IEEE Workshop Comput. Intell. Biometrics Identity Manage. (CIBIM)*, Apr. 2013, pp. 14–21.

[10] J. M. Smereka, V. N. Boddeti, and B. V. K. V. Kumar, "Probabilistic deformation models for challenging periocular image verification," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 9, pp. 1875–1890, Sep. 2015.

[11] F. Juefei-Xu, K. Luu, M. Savvides, T. D. Bui, and C. Y. Suen, "Investigating age invariant face recognition based on periocular biometrics," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, Oct. 2011, pp. 1–7.

[12] E. DeLong, D. DeLong, and D. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[13] D. L. Woodard, S. Pundlik, P. Miller, R. Jillela, and A. Ross, "On the fusion of periocular and iris biometrics in non-ideal imagery," in *Proc. IEEE Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2010, pp. 201–204.

[14] A. Kumar, "Neural network based detection of local textile defects," *Pattern Recognit.*, vol. 36, pp. 1645–1659, Jul. 2003.

[15] R. Jillela and A. Ross, "Mitigating effects of plastic surgery: Fusing face and ocular biometrics," in *Proc. IEEE 5th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2012, pp. 402–411.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[17] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[18] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1891–1898.

[19] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1701–1708.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] X. Li and Z. Cui, "Deep residual networks for plankton classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–4.

[22] S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh, "Periocular biometrics: When iris recognition fails," in *Proc. 4th IEEE Int. Conf. Biometrics, Theory Appl. Syst. (BTAS)*, Sep. 2010, pp. 1–6.

[23] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[24] L. Wan, M. Zeiler, S. Zhang, Y. Lecun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1058–1066.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[26] C. N. Padole and H. Proenca, "Periocular recognition: Analysis of performance degradation factors," in *Proc. IEEE 5th IAPR Int. Conf. Biometrics (ICB)*, Apr. 2012, pp. 439–445.

[27] *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)*. Accessed on Mar. 2016. [Online]. Available: http://www.image-net.org/challenges/LSVRC/

[28] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, Mar. 2016. [Online]. Available: http://vis-www.cs.umass.edu/lfw/

[29] H. Proença and L. A. Alexandre, "The NICE.I: Noisy iris challenge evaluation—Part I," in *Proc. 1st IEEE Int. Conf. Biometrics, Theory Appl. Syst. (BTAS)*, Sep. 2007, pp. 1–4.

[30] (2016). *FRGC Dataset*, accessed on Mar. 29, 2016. [Online]. Available: http://www.nist.gov/itl/iad/ig/frgc.cfm

[31] G. Bradski, "The OpenCV library," *Dr. Dobb's J. Softw. Tools*, vol. 25, pp. 120–126, Nov. 2000.

[32] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Dec. 2001, pp. I-511–I-518.

[33] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Computer Vision*. Berlin, Germany: Springer, 2012, pp. 566–579.

[34] (2016). *FOCS Dataset*, accessed on Mar. 29, 2016. [Online]. Available: http://www.nist.gov/itl/iad/ig/focs.cfm

[35] (2016). *CASIA.V4 Dataset*, accessed on Mar. 29, 2016. [Online]. Available: http://biometrics.idealtest.org/dbDetailForUser.do?id=4

[36] (2016). *UBIpr Dataset*, accessed on Mar. 29, 2016. [Online]. Available: http://socia-lab.di.ubi.pt/~ubipr/

[37] (2016). *CIFAR-10 Dataset*, accessed on Mar. 29, 2016. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html

[38] *Cuda-Convnet—High-Performance C++/CUDA Implementation of Convolutional Neural Networks—Google Project Hosting, Code.google.com*, accessed on Mar. 29, 2016. [Online]. Available: https://code.google.com/p/cuda-convnet/

[39] *Caffe | CIFAR-10 Tutorial, Caffe.berkeleyvision.org*, accessed on Mar. 29, 2016. [Online]. Available: http://caffe.berkeleyvision.org/gathered/examples/cifar10.html

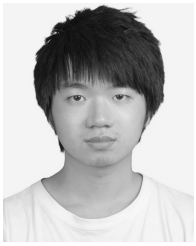
[40] (Nov. 2016). *Weblink to Download Implementation Codes to Reproduce Results*. [Online]. Available: http://www.comp.polyu.edu.hk/~csajaykr/scnn.rar

[41] *Biometric Evaluations Homepage, Nist.gov*, accessed on Mar. 29, 2016. [Online]. Available: http://www.nist.gov/itl/iad/ig/biometric_evaluations.cfm

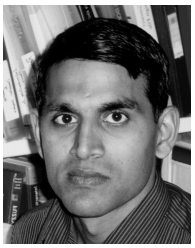
[42] Alamy. *Stock Photo—Epa01034158 Demonstrators Cover Their Faces During Clashes at the End of the Far Leftist 'No Bush-No War' Rally*, accessed on May 30, 2016. [Online]. Available: http://www.alamy.com/stock-photo-epa01034158-demonstrators-cover-their-faces-during-clashes-at-the-97583185.html

[43] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2892–2900.

[44] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.



Zijing Zhao (S'15) received the B.E. degree in software engineering from Zhejiang University in 2012, and the M.Sc. degree (Hons.) in software technology from The Hong Kong Polytechnic University in 2013, where he is currently pursuing the Ph.D. degree with the Department of Computing. His research interests are in computer vision, pattern recognition, and machine learning. He received *The Hong Kong Ph.D. Fellowship Award* in 2014.



Ajay Kumar (S'00–M'01–SM'07) received the Ph.D. degree from The University of Hong Kong, Hong Kong, in 2001. He was an Assistant Professor with the Department of Electrical Engineering, IIT Delhi, New Delhi, India, from 2005 to 2007. He is currently an Associate Professor with the Department of Computing, The Hong Kong Polytechnic University. He holds five U.S. patents, and has authored on biometrics and computer vision based industrial inspection. His current research interests are on biometrics with an emphasis on hand biometrics, vascular biometrics, iris, and multimodal biometrics. He was on the Editorial Board of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY from 2010 to 2013, and served on the program committees of several international conferences and workshops in the field of his research interest. He was the Program Chair of ICEB, Hong Kong, in 2010, and a Program Co-Chair of IJCB, Washington, DC, USA, in 2011, the ICB Madrid, in 2013, and the CVPR 2013-2016 Biometrics Workshops. He has also served as the General Co-Chair of the IJCB in 2014, Tampa, and the ISBA in 2015, Hong Kong, and delivered the first tutorial on contactless 3-D fingerprint identification during the CVPR 2015 held in Boston, USA. He currently serves as an Area Editor of *Pattern Recognition Letters* and served on the IEEE Biometrics Council as the Vice President for Publications from 2011 to 2015.