Pattern Recognition Letters
journal homepage: www.elsevier.com

Convolutional Neural Networks for Ocular Smartphone-Based Biometrics

Karan Ahuja[a], Rahul Islam[a], Ferdous A. Barbhuiya[a], Kuntal Dey[b,**]

[a]*Indian Institute of Information Technology, Guwahati, India*
[b]*IBM Research, New Delhi, India*

ABSTRACT

Ocular biometrics in the visible spectrum has emerged as an area of significant research activity. In this paper, we propose a hybrid convolution-based model, for verifying a pair of periocular images containing the iris. We compose the hybrid model as a combination of a supervised and an unsupervised convolution, and augment with the well-known geometry-based Root SIFT model. We also compare the performance of two convolution-based models against each other, as well as, with the baseline Root SIFT. In the first (unsupervised w.r.t target dataset) convolution based deep learning approach, we use a stacked convolutional architecture, using external models learned *a-priori* on external facial and periocular data, on top of the baseline Root SIFT model applied on the provided data, and apply different score fusion models. In the second (supervised w.r.t target dataset) approach, we again use a stacked convolution architecture; but here, we learn the feature vector in a supervised manner. On the MICHE-II dataset, we obtain an AUROC of 0.946 and 0.981, and EER of 0.092 and 0.066, for the two models respectively. The hybrid model we propose, which combines these two convolutional neural networks, outperforms the constituents, in case the both the images arise from the same device type, but not necessarily so otherwise, obtaining a AUROC of 0.986 and EER of 0.053. We also benchmark our performance on the standard VISOB database, where we outperform the state of the art methods, achieving a TPR of 99.5% at a FPR of 0.001%. Given the robustness and significant performance our methodology, our system can be used in real-life applications with minimal error.

1. Introduction

Identifying individuals as genuine versus impostors, using facial features such as matching of the iris, periocular region and face, have emerged as areas of research interest. As cameras are now becoming ubiquitous in the Internet of things, their use in the field of biometric based access control applications is ever increasing, as showcased in the product HYPR [1] and paper by [2]. Therefore, there is a need for robust, reliable and accurate smartphone based biometric authentication under constrained scenarios. In this paper, following the definition given by [3], the term periocular means the area surrounding the eye as well as the eye - i.e. containing the iris and sclera also. The qualitative problem at hand is of verifying whether a pair of periocular images taken in the visible spectrum belong to the same person or not. Computationally, this problem manifests as measuring the similarity between a pair of periocular region images, using a combination of features derived from the iris region and the surrounding periocular region. The aim is to verify whether a given pair of images are the same or not, optionally classifying the identity of the subject. The applications of such systems are manifold, such as performing multimodal authentication on smartphones discussed by [4].

Over time, computer vision and image processing based techniques have significantly matured. With mobile phone cameras becoming a *de facto* practice for clicking photographs, CCTV-based surveillance systems gaining prominence, the cost of digital photographs going down to practically negligible, and the quality of photos improving significantly with advancement of hardware, using computer vision based techniques for biometrics is also becoming more prevalent than ever before. These advancements have also created a different avenue to solve the problem at our hands.

Several novel approaches have been attempted towards ocu-

**Corresponding author: Tel.: +91-987-138-8275;
e-mail: kuntadey@in.ibm.com (Kuntal Dey)

lar and periocular biometrics. Early works, such as [5], [3] and [6], explore the feasibility and use of periocular biometrics. A number of works were presented in the recent periocular identification challenge by [7], that provided a new database, namely the VISOB database. It was observed that deep learning approaches, such as the works by [8] and [9], were more effective compared to the others such as [10]. Works using iris segmentation have been conducted in different studies, such as by [11], [12] and [4]. A recent challenge was organized as MICHE [13] for iris-based (ocular) biometrics.

In this paper, we propose a hybrid model, that uses two independent underlying models, to solve the problem at hand. A preliminary version of our work had appeared in ICPR 2016 [14]. We note that, with deep learning systems such as deep convolutional neural networks (CNNs - also known as ConvNets), it is easy to inspire from the transfer learning paradigm proposed by [15], and apply a fusion of feature representations learned externally as well as from the training data at hand. Since, multiple databases, such as VISOB and MICHE-II, provide photographs of the ocular and periocular regions, we propose a CNN based learning for each, leading to two independent CNN based models. We further note that, the recent work on Openface by [16], has given a method to create a 128-dimensional feature vector for face images for the purpose of face verification, and empirically observe that this method works well for the task of verification from partial face images as well. We, thus, create our proposed hybrid model, as a fusion of the two independent CNN models, the features given by OpenFace and Root SIFT.

In order to test the performance of the proposed hybrid model on the MICHE-II dataset, we compare it with a geometry-based baseline approach, as well as the two underlying CNN based approaches that constitute the hybrid model. For geometry-based baseline, we use the well-known Root SIFT method [17]. Root SIFT calculates the image descriptor of the iris images segmented out of the images of the MICHE-II dataset, and subsequently matches an image with the other using a k-nearest neighbor approach.

Both the deep learning models are based on CNNs, and both the models use a stacked layer architecture. In one CNN-based model, we follow a unsupervised approach, using external *a-priori* knowledge along with Transfer Learning techniques. We train on the VISOB dataset, and obtain a $1,024$-sized feature vector of the periocular region. We also benchmark this approach on the VISOB validation dataset. As mentioned earlier, we use the 128-dimensional facial feature vector given by OpenFace too. We use these two feature vectors, with Root SIFT, and combine the scores assigned by each of these three subsystems, to calculate a dissimilarity score, using simple averaging as well as linear regression based techniques. Thus, with respect to the provided MICHE-II database, this model is unsupervised in nature.

In the second CNN-based baseline approach, we avoid using external *a-priori* knowledge, and solely rely upon the provided MICHE-II dataset to perform CNN-based deep neural network learning. This approach is supervised by nature. We pass each training image through a ConvNet, and subsequently create a 512-sized feature vector for each training image. For each test image, we construct its 512-feature vector, and compare this vector with each of the training vectors using cosine similarity, to find the best-match image.

Note that, in both CNN-driven baseline models, as well as in our final hybrid model, we generate data using known augmentation techniques, to further improve the performance of our system. On the VISOB database, we achieve an EER of 0.0059%, with a a TPR of 99.5% at a FPR of 0.001%. On the MICHE-II test dataset, we obtain an AUROC of 0.946 and 0.981, and EER of 0.092 and 0.066, for the two CNN-based baseline models respectively, when testing under the same device constraint. In the hybrid model, we obtain an EER of 0.057 and AUROC of 0.985. Clearly, the hybrid model, which is essentially a composition of the two CNN-based baseline models, outperforms all the other models on the MICHE-II dataset, and delivers the most optimal performance under the same device constraint. The high performances that our model yields is encouraging. We note that, both the baseline CNN models, as well as our final hybrid model, are practically reasonable candidates for deploying in real-life applications.

Thus, the contributions of our work are as follows.

- We propose two indepedent CNN-based models and a hybrid model to solve the periocular biometric based verification problem.
- We create a first CNN-based unsupervised model, that leverages the benefits of transfer learning, using the VISOB database and OpenFace facial feature identification system, and combining that with Root SIFT.
- We create a second CNN-based supervised model, that uses only the images from the provided MICHE-II dataset, and uses a cosine similarity metric on the derived feature vector for measuring similarity between image pairs.
- We propose a deep CNN based hybrid model, as a fusion of (a) the two independent CNN based models, (b) the set of features given by OpenFace and (c) Root SIFT.
- We provide an empirical comparison of the two CNN baselines between each other, as well as with a baseline Root SIFT model. We observe the second (supervised) CNN based model to outperform the first. Both the models significantly outperform the Root SIFT model.
- We examine the performance of our hybrid model, with respect to three baseline models on standard databases, namely Root SIFT, and the two CNN based models.
- The proposed hybrid model, outperforms the two independent CNN based models as well as the baseline Root SIFT model. We validate this by testing in settings where the training and test images stem from the same device type.

The rest of the paper is as follows. The literature is covered in Section 2. The details of our methodology, including the design principles and the models, are presented in Section 3. Section 4 explores the outcome of applying our methodology on the target dataset. Finally, we provide a brief discussion in Section 5 and conclude in Section 6.

2. Related Work

As the computer vision and image processing techniques have matured with time, a number of novel approaches towards the ocular and periocular biometrics problem (and closely related problems) have been recently introduced by several researchers. Some of these works have attempted to address the problem by investigating the iris, while others have attempted to inspect the periocular regions also, that is, the regions that also surround the eye. Recently [18] reviewed the research progress in the area and discussed existing algorithms and the limitations of each of the biometric traits and information fusion approaches.

Early works, such as [5], [19] and [3], establish the feasibility of using periocular images for biometric identification. They use texture and point operators to extract global and local information from the periocular region, and use these features to represent and match the region. They study the impact of several factors in periocular verification, such as the effectiveness of incorporating and disguising eyebrows in the feature set, the effect of masking the iris and the eye region, the effects of pose validation and occlusion, and the effectiveness of using a fusion of face and periocular biometrics.

[6] evaluate the utility of the periocular region appearance cues for biometric identification. They demonstrate the effectiveness of periocular biometrics to be at par with face-based recognition. They divide the periocular image into salient patches using local appearance based features, and compute histograms of texture and color from each patch. They match the images by computing distances between the features.

In a recent periocular identification challenge [7], a database with the title VISOB was provided, and a number of approaches of identifying individuals were presented, where the images were collected in the visible spectra under different lighting conditions, namely daylight, dim light and office lighting, and different devices, namely Oppo mobile phones, Samsung mobile phones and Apple iPhones. Several interesting works emerged in the challenge.

With the VISOB dataset, [8] propose to extract texture features from periocular images using maximum response filters, and subsequently classify these features using deeply coupled auto-encoders. They model with a 4-layer deep auto-encoder for performing unsupervised feature learning, and finally perform a supervised softmax based verification. [9] propose a framework, based upon collaboratively represented features from deep sparse filtering. Some other works, with lesser yields, also have been proposed, such as the 2-phase approach by [10] that uses a multinomial Bayesian Learning followed by Dense SIFT. All these works attempt to use the minimum distance between the enrollment class and probe class, and use this distance to assign a classification label of genuine versus impostor.

Pertaining to the task of iris segmentation, [12] design an unsupervised iris defects detection method based on the underlying multispectral spatial probabilistic iris textural model. They perform adaptive thresholding, that would be effective for high resolution mobile device measurements, in the visible and near-infrared spectrum. Their model is based upon adaptive parameter learning for iris texture, and checking for iris reflections using recursive prediction analysis techniques. For iris recognition in the visible spectrum, [11] describes an integrated scheme for noisy iris recognition in adverse conditions. They perform iris matching by combining local features, such as linear binary patterns (LBPs) and discriminable textons (BLOBs). They refine their techniques ad hoc, to keep their approach amenable to work with images captures in varying visible light conditions, as well as noises arising from distance, hardware limitations such as resolution, and scarce user collaboration such as blurring, off-axis iris, and occlusions by eyelids and eyelashes.

For mobile devices in particular, [4] propose a system, that combines the recognition of user's iris and a image forensic field technique for camera source identification, namely sensor pattern noise of user devices, for authentication of users. They perform fusion of the multimodal inputs at two levels - the feature level and at score level. For feature level fusion, concatenate the feature vectors obtained from their sensor recognition and iris recognition modules, and then perform feature selection. For score level fusion, they compute the distance matrices for the two recognition modules, and apply different score normalization techniques. [20] also propose a multimodal recognition. They perform fusion of face and iris features, for the purpose of recognition.

A recent challenge was also organized as MICHE-I [13] for iris-based (ocular) biometrics, and a subsequent follow-up organized as part the ICPR 2016 Conference presented in [21]. The current paper uses the MICHE-II data released as part of the this challenge. The novelty in our work, in context of the prior literature, is in the proposed hybrid CNN model that (a) encompasses (i) unsupervised transfer learning using *a-priori* knowledge present in the externally created related CNN feature vectors, and (ii) supervised feature vector learning on the provided MICHE data training set, (b) augments that with the traditional geometry-based Root SIFT model, and (c) an augmentation of this by a benchmarking across the two state-of-the-art databases, namely the VISOB and the MICHE-II databases.

3. Methodology

For the purpose of verifying individuals, we explore the following algorithms encompassing the task of iris and periocular verification. We propose a baseline Root SIFT method for iris verification, a Deep Learning based model for periocular verification and two further subsequent models that learn discriminant appearance based features from the periocular region for verification. Building on the insights of state of the art facial recognition models such as [22], it is seen that supervised methods have a clear advantage over unsupervised ones. As showcased in [15], training and testing across different distributions can degrade performance considerably. However, fitting a model to a small dataset decreases it's robustness and generalization to other datasets. We therefore investigate both these methods - learning a supervised and unsupervised verification metric. We test our results on the VISOB and MICHE-II databases.

In the first model, we aim at learning an unsupervised metric, that generalizes well across several datasets. In this setting, no training whatsoever is performed on the MICHE-II database. In the second model, we employ a supervised learning paradigm that learns feature representation for comparison and verification on the MICHE-II dataset.

3.1. Baseline Model

Inspired by SIFT based models for ocular biometrics in the visible spectrum such as [7], [10] and [23], we make use of Dense SIFT keypoints for matching irises. Iris verification is comprised of two main stages: iris segmentation and feature matching. For the first task, the iris is extracted out of the image using the segmentation algorithm described by [12]. The algorithm provides us with the segmented and normalized iris image of dimensions 600×100 pixels along with a defects mask. We first overlay the segmented iris image with the binary mask to get the iris image rid of any occlusions. We then compute Dense color Root SIFT descriptors from the RGB channels, as proposed by [17], which gives us keypoints with identical size and orientation. The advantage of Root SIFT over traditional SIFT given by [24] is that it employs a Hellinger kernel instead of the standard Euclidean distance to measure the similarity between SIFT descriptors. Matching between descriptors is performed by comparing each local extrema using a nearest neighbor matcher given by [25]. The dissimilarity score d is defined as:

$$d = \left(1 - \frac{|Matches|}{min(|KeyPts\_img1|, |KeyPts\_img2|)}\right) \quad (1)$$

In equation 1, |KeyPts| describes the number of SIFT keypoints detected in the respective images, and |Matches| defines the number of keypoint matches returned by the nearest neighbor matcher.

3.2. VisobNet

Deep learning systems have achieved state of the art accuracies in face recognition tasks. However, they require large a large training database to learn their models. Alternatively, the use of transfer learning described by [15] is often used to solve this problem. Here the feature representation is learned on an external dataset. Motivated by the success of such approaches, we employ a similar approach in which we train our model on a relatively larger database and test on a more challenging one. The model automatically learns appearance-based features by using a Deep Convolutional Neural network. We train our CNN on a multi-class recognition task, namely to classify the identity of the periocular image. The overall architecture is depicted in Table 1.

First, the periocular region is extracted from the given image by creating a rough bounding box around the eye, the dimension of which are given as a function of the iris center and radius returned by the unsupervised iris segmentation proposed by [12]. This RGB (3 channel) periocular image is re-sized to 32 pixels × 48 pixels and given as input to the convolution layer 1. We use a convolution kernel of size 3×3 in all the convolution layers and max-pooling after two consecutive convolution layers. This makes the output of the convolution network robust to small errors and translations. We also use dropout layers for regularization to encourage sparsity and prevent over-fitting.

Finally, the last two layers of the network are fully connected to capture the correlation between the features captured in different parts of the eye. The dense layers in Table 1 represent the fully connected layers. We use ReLU proposed by [26] as the activation function except in the last layer where we use a softmax classifier, which produces a distribution over the class labels. The goal of training is to maximize the probability of the correct label. This is achieved by minimizing the cross-entropy loss for each training sample. We train the CNN using Stochastic Gradient Descent (SGD) [27] with standard back-propagation and Momentum (set to 0.9) [28]. We train the model with a learning rate of 0.01 for all layers, and a batch size of 256 for $1,500$ epochs. We also employ real-time data augmentation to increase the samples for training. We use the Keras library developed by [29] for training our model.

We take the output of the Fully Connected layer 1 to get the $1,024$-dimensional feature vector of the periocular image. This representation is in contrast to traditional representations proposed in literature that normally pool descriptors, and use it as input to a classifier. We investigate the use of both a supervised and unsupervised verification metric to test the robustness of our ConvNet. For our unsupervised similarity metric, we take the cosine similarity between two feature vectors of the periocular images. For the supervised metric, the probability of classification of the probe class to the enrollment class is used to classify the image and assign it a similarity (or dissimilarity) score.

3.3. Model 1

Figure 1 illustrates our proposed framework for Model 1. The model consists of three integral parts for the purpose of verification, namely OpenFace, Visobnet and RootSIFT, as described below.

3.3.1. OpenFace

OpenFace proposed by [16] is a general purpose face recognition library that is well-suited for mobile scenarios. Given a facial image, it outputs a 128-dimensional feature vector of that image. Although, it is crafted for face verification, we find it to perform well for the task of verification from partial face images as well. OpenFace takes two facial images (or partial face images) as inputs and subsequently outputs the predicted similarity score of two images by computing the squared L2 distance between their representations. Since the representations are on the unit hypersphere, the scores range from 0 to 4.0. We then normalize the score to a range from 0 to 1 to get the dissimilarity score. As we do not tune OpenFace for our target database, it helps us understand to what extent can existing state-of-the art methods for face verification be employed for the task of ocular biometrics.

3.3.2. VisobNet

This is the ConvNet described in Section 3.2. We use it as a part of Model 1 in an unsupervised manner, that is, we calculate

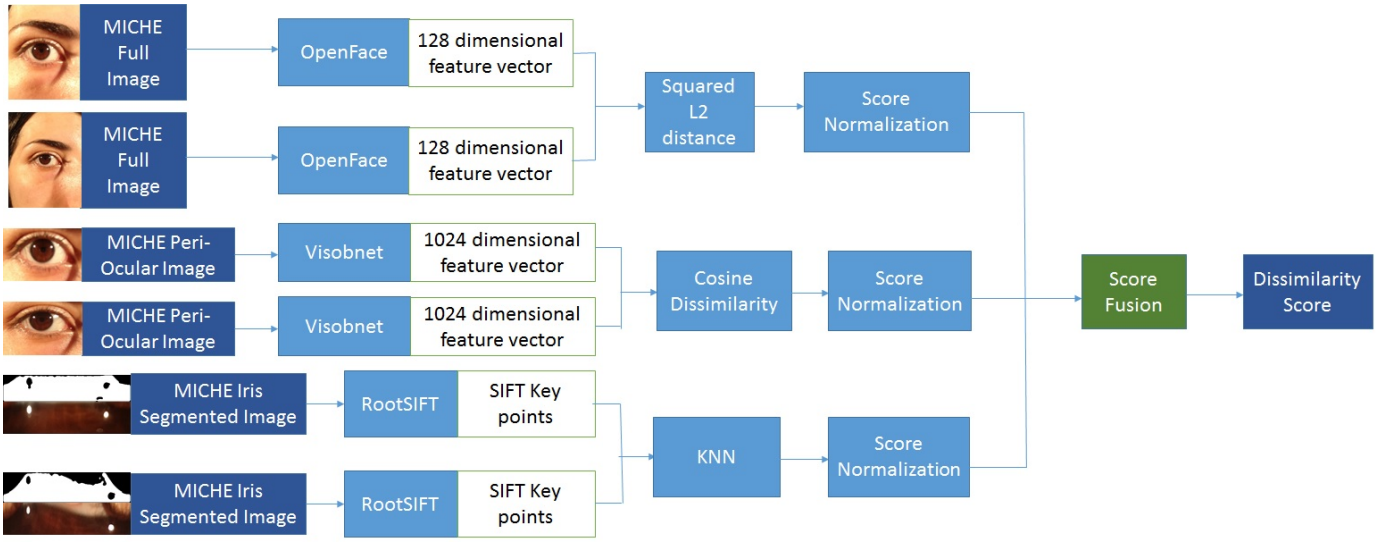

Fig. 1. Schematic representation of proposed Model 1.

the cosine similarity between the pair of output feature representations obtained by passing the periocular image though the feed forward network.

3.3.3. RootSIFT

This is the Root SIFT based baseline model, described in Section 3.1. The dissimilarity score for a given iris pair is used for verification purposes.

3.3.4. Score Fusion

We first normalize the scores to bring them within the fixed numerical range of [0,1]. We then employ two approaches for score fusion. In the first approach we simply take an average of all the scores. Hence, the model and dissimilarity metric remains completely unsupervised. In the second approach we train a linear regressor on 5% of the total image pairs of the target database. Note that, this is a supervised approach, carried out to compare its results with the unsupervised verification metric.

3.4. Model 2

In this model we employ a supervised CNN to learn discriminative feature representations from the target dataset, rather than opting for transfer learning based approaches. As noted earlier, in the domain of face verification and other recognition tasks, supervised methods tend to show a clear advantage over unsupervised ones. We therefore, employ this model with the hope of contrasting it's performance with our proposed unsupervised Model 1. The details of our CNN Model are captured in Table 2. The advantage of employing appearance based Convolution Neural Network is that it is able to visualize the iris from periocular region on the fly without prior need for segmentation. As our target dataset contains only a few over $3,000$ images, we resort to data augmentation to increase the robustness and generality of our model. For data augmentation, we rotate the image between 0 to 30 degrees, randomly shift the images horizontally and vertically by 0.1 of their total width and height respectively. We also flip the images horizontally and also zoom it in between 0.7 to 1.3 times it's original size. The CNN model details are similar to Section 3.2 in terms of learning algorithm, rates and kernel sizes, with the only difference being that we train this model for a $1,000$ epochs because it is shallower in comparison and hence converges faster. The input of the model is a re-sized RGB image (we pass the whole image without prior segmentation) having dimensions of 64×96, and it's output is a 512 dimensional feature vector. Similar to Section 3.2 we employ a cosine similarity to get the similarity between two feature vectors.

3.5. Hybrid Model

This model is a amalgamation of our unsupervised and supervised model. Here the score of Model 2 is used in conjunction with the score of Model 1 (average of all scores in Model 1) to compute a fused dissimilarity score. The fusion is computed as the average of the two scores from each model. We choose this approach due to its generality and simplicity.

4. Evaluation

We take the MICHE-II dataset [13] as the target database for evaluating the performance of the baseline RootSIFT model and the two proposed CNN-based models. We also evaluate the accuracy of VisobNet on the VISOB database [7]. In this section, we first describe the datasets used for evaluation of different models, then we present our detailed evaluation and comparison with state-of-the art systems, followed by an analysis of our results.

For experiments, we use a hardware configuration of Intel Pentium CPU 2020M @ 2.40 GHz and 4 GB RAM. Our methodology for Model 1 achieves an execution time of approximately 1.7 seconds for inference and 130 seconds to run the externally provided segmentation method for a given image pair. As our Model 2 does not require any prior segmentation, it achieves a smaller execution time of 0.6 seconds for verifying a given image pair.

4.1. Data Description

We evaluate the performance of our system on the VISOB and MICHE-II datasets.

The MICHE-II dataset is an iris biometric dataset captured under uncontrolled settings using mobile devices in the visible spectrum. Figure 2 depicts sample images taken from the database. It is captured under the same paradigm as MICHE-I with respect to the environment, mode of capture, *etc.* Its training dataset comprises of over $3,000$ images, across all environments (across 2 different lighting conditions), devices (across 3 different devices) and eyes (left/right), and has 75 distinct labels (unique subjects), while its test dataset comprises of 120 images of the left and the right eyes combined captured using two devices, namely Samsung Galaxy S4 and Apple iPhone 5. While some of the subjects present in it are part of the MICHE-II training database, most of its subjects are new and are not a part of the training dataset.

The VISOB dataset is a large scale database to test the performance of mobile ocular biometric schemes in visible spectrum. Figure 3 showcases sample images taken from the database. It consists of periocular eye images from 550 healthy adult volunteers using three different phones (Samsung, IPhone and Oppo) and three different lighting conditions (Office, Day and Dim Light). It captures the images across two Visits. We concern ourselves with the publicly available Visit 1 dataset, which was provided to the contestants of the ICIP 2016 VISOB Challenge for evaluation and reporting. The Visit 1 database is a closed system having a total of 48250 enrollment images and 46797 verification images. All the images provided in VISOB are preprocessed and cropped to retain only the periocular region of size 240×160 using a Viola-Jones based eye detector discussed in [7].

Due to the relatively larger size and preprocessed periocular images, VISOB makes a good training dataset for CNN's (also providing a very large verification database for testing). However, MICHE-II test-dataset introduces a more challenging, open, unconstrained yet smaller dataset making it ideal as our target database for testing. Therefore, we train the proposed VisobNet on the enrollment database of VISOB and test it's performance as a supervised metric on the VISOB verification database, and as an unsupervised verification metric with respect to MICHE-II test dataset. As noted by [9], VISOB database fails to provide fixed number of enrollment samples at a per device and per environment setting. We therefore, train on all the enrollment images together, regardless of their environment or device of capture. Therefore, our deep learning network is able to leverage better hierarchical and discriminative features from a much more vast and diverse enrollment dataset.

For Model 2, we train it's ConvNet as a recognition system on 80% of the MICHE-II training dataset and use the remaining 20% for validation. For testing the Baseline Root SIFT model, Model 1 and Model 2, we use MICHE-II test dataset as the target database for evaluation, for reasons discussed above.

4.2. Model Evaluation

We present the results of the proposed schemes on the VISOB and MICHE-II test dataset. The results are represented using Receiver Operating Curves (ROC) and Equal Error Rates.

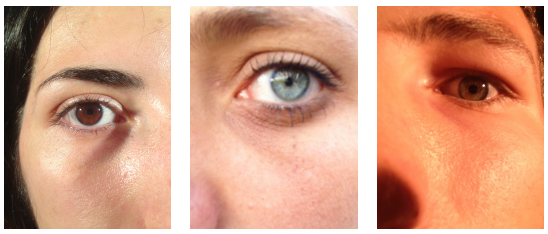

Fig. 2. Sample Images from the MICHE-II Database

4.2.1. Evaluating on the VISOB dataset

We present the results of VisobNet for periocular recognition on the VISOB database. Similar to [8], experiments are carried out by training the CNN on the enrollment database and testing on the validation (verification) database. Since, the VISOB database is a closed ID system, we take the probability of the probe class to the enrollment class to classify the images and assign it a similarity (or dissimilarity) score.

We evaluate it's performance of periocular recognition by calculating it's rank one accuracies across different devices and lighting condition, as depicted in Table 3. Figure 12 depicts the Cumulative Match curves of the VISOB verification database. Our ConvNet achieves a rank one accuracy of 93.49% on the VISOB database, which is much higher than the average and best case accuracy of 63.98% and 79.49% reported by [10] on the same database.

For the task of verification, we output the probability distributions of a probe class across all the enrollment classes. We also compare the algorithms performance with sate-of-the art algorithms proposed by [8] and [9] in Table 4. The ROC performance of the proposed method for the images captured across three different devices -Samsung, Oppo and I-Phone, each with three different lighting conditions - Day, Dim and Office, for each eye - Left and Right, is depicted in Figures 13, 14 and 15 respectively. It can be observed that our system beats the current state-of-the art by a considerable margin. Across all devices and environments, we achieve a True Positive Rate (TPR) of 99.5% at a False Positive Rate (FPR) of 0.001%. We clearly out-perform the best performance of the front-runners of the ICIP VISOB challenge [7]. [9] and [8], who achieve a best case accuracy of 97.56% and 93.98%. Our system achieves a considerably low Equal Error Rate (EER) of 0.0059%, which goes on to prove it's robustness and reliability.

The considerably high accuracy achieved by VisobNet on the VISOB verification database can be attributed to the following:

- All the images are cropped and preprocessed therefore aligning all the periocular images and making feature representation easier.
- The VISOB database is a closed ID system, that is, the probe is always one of the known identities in the database.
- The fact that we train the ConvNet from all the enrollment images across different devices and environments makes the feature representation richer and more robust.
- The biggest contributor to the very high True positive Rate at a low False Acceptance Rate, is the use of the softmax

Fig. 3. Sample Images from the VISOB Database

function that outputs the normalized probabilities for each class, via a cross-entropy loss function that minimizes the negative log likelihood of the correct class. Also the use of ReLU as the activation function, makes the ConvNet produce highly non-linear and sparse features.

The advantages of using a supervised verification metric as described above is that it achieves a very high performance across devices and environments, without the need for customizing features for each scenario. However, the given metric may degrade in an Open World system, or systems where there are considerably less samples for a given class. Also, the addition of new classes to the network would require to fine-tune the parameters to incorporate them. Therefore, we also explore an unsupervised verification metric, in the form of inner product between two feature representations, to evaluate the generality of VisobNet. For this purpose, we test it one the MICHE-II test dataset, the results of which are given in Section 4.2.2.

4.2.2. Evaluation on MICHE-II dataset

We evaluate the performance of the baseline Root SIFT method, OpenFace, VisobNet, Model 1, Model 2 and the Hybrid Model on the MICHE-II test dataset. A test verification process is carried out, by comparing each test dataset image with one another, in all possible combinations, under each of the above settings. We perform empirical evaluation of our models under the following paradigms.

Same-Eye versus Cross-Eye: Under the *same-eye paradigm*, we hypothesize that the left and right iris of a given person are different from each other. Hence, we compare the Left Eye Images with Left Eye Images and Right Eye Images with Right Eye Images. Under the *cross-eye paradigm*, we ignore the possibility that left and right eyes could produce different features, and merge all the eye images for the comparison. The rationale behind making this apparently counter-intuitive assumption are to exploit the following. (a) Data augmentation with horizontal flip: In the data augmentation process during the deep CNN training, we also perform horizontal flip of the images, thereby the left and right eyes also getting "interchanged" in the learning process. (b) Feature similarity: In the given image dataset features, only minor dissimilarities exist between left and right eye images of most of the given persons. We observe similar performances in these two paradigms.

Same-Device versus Cross-Device: Under the *same-device paradigm*, we compare images taken from the same device type with each other. That is, we compare images taken from Samsung Galaxy S4 only with other images taken from Samsung Galaxy S4, and images taken from Apple iPhone 5 only with other images taken from Apple iPhone 5. Under the *cross-device paradigm*, we compare between the images agnostic of the device type from which any of the images were taken from. Note that, we experiment with both the *same-eye (SE)* and *cross-eye (CE)* with the *same-device (SD)* and *cross-device (CD)* paradigms, and observe similar performance outcomes between same-eye and cross-eye testing, whereas there is a stark improvement in results when migrating from cross-device to same-device paradigm. This can be seen in Tables 5 and 6 which correspond to EER and AUROC respectively for the various methods. Here, Model 1 LR refers to the Linear Regression based supervised score fusion technique, as opposed to EQ which refers to the unsupervised average based score fusion. Figures 4, 5, 6, 7, 8, 9 and 10 showcase the ROC curves of the various methods discussed in Section 3. In these figures, the label *Default* corresponds to the CD_SE paradigm, and *Same Device* corresponds to the SD_SE paradigm. For our hybrid model we achieve an EER of 0.352 and 0.057, and AUROC of 0.736 and 0.985 in the CD_SE and SD_SE paradigms respectively. The FAR-FRR Curve for this can be found in Figure 11. Thus, the best performance of our proposed scheme is achieved by the hybrid model on the SD_SE paradigm. It should also be noted, that the hybrid model showcases similar performances for both left and right periocular images.

5. Discussion

As shown in Section 4, our system delivers a stark improvement over the baseline approach and current state of the art algorithms. This can be attributed to the use of deep learning neural network models. While supervised models clearly outperform the unsupervised ones, it is interesting to note that unsupervised models learned on different (external) datasets also provide reasonable accuracy, when applied on the current dataset. One interesting observation is that OpenFace, a model created for facial recognition, performs reasonably well on the MICHE-II test database, where only partial faces are visible. The success of such models, opens further avenues such as using of pre-trained models - trained on a larger (external) dataset, albeit for a slightly different task such as face or periocular recognition, and fine-tuning its last layers for addressing the complexities of the target dataset; thus maintaining the generality and robustness of the system, and at the same time fitting the model better for the target dataset. It will also be of interest to explore feature embeddings that directly correspond to image similarity, such as the Weighted X^2 distance.

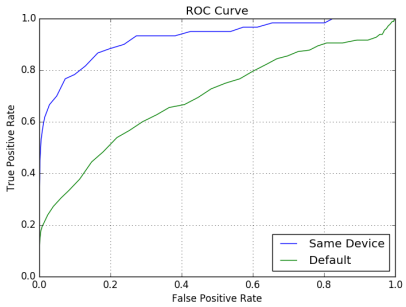

Fig. 4. OpenFace ROC on MICHE-II

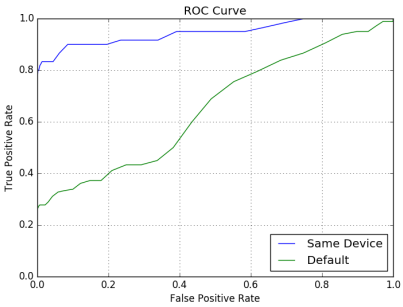

Fig. 5. Model 1 EQ ROC on MICHE-II

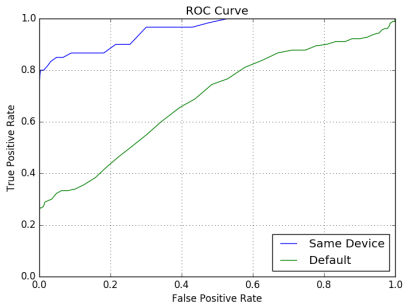

Fig. 6. Model 1 LR ROC on MICHE-II

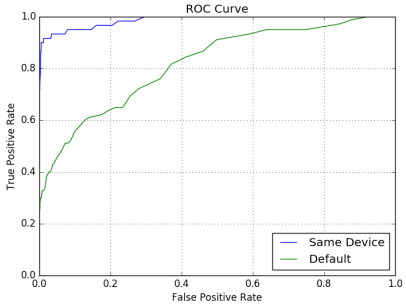

Fig. 7. Model 2 ROC on MICHE-II

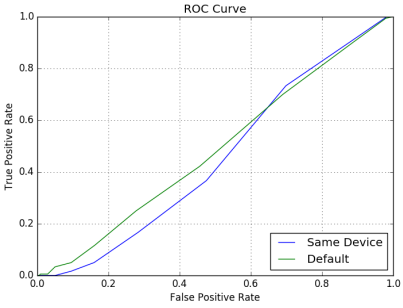

Fig. 8. Root SIFT ROC on MICHE-II

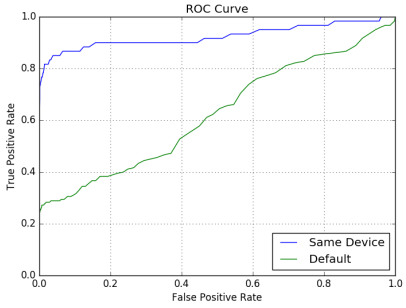

Fig. 9. Visobnet ROC on MICHE-II

Table 1. Our architecture for Visobnet features. The output size is given by filters×rows×cols.

Layer	Output Shape	Params
convolution2d_1	(32, 32, 48)	896
activation_1	(32, 32, 48)	0
convolution2d_2	(32, 30, 46)	9248
activation_2	(32, 30, 46)	0
maxpooling2d_1	(32, 15, 23)	0
dropout_1	(32, 15, 23)	0
convolution2d_3	(64, 15, 23)	18496
activation_3	(64, 15, 23)	0
convolution2d_4	(64, 13, 21)	36928
activation_4	(64, 13, 21)	0
maxpooling2d_2	(64, 6, 10)	0
dropout_2	(64, 6, 10)	0
convolution2d_5	(128, 6, 10)	73856
activation_5	(128, 6, 10)	0
convolution2d_6	(128, 4, 8)	147584
activation_6	(128, 4, 8)	0
maxpooling2d_3	(128, 2, 4)	0
dropout_3	(128, 2, 4)	0
flatten_1	(1024)	0
dense_1	(1024)	1049600
activation_7	(1024)	0
dropout_4	(1024)	0
dense_2	(586)	600650
activation_8	(586)	0
	Total params	1937258

Table 2. Our architecture for Model 2-based CNN. The output size is given by filters×rows×cols.

Layer	Output Shape	Params
convolution2d_1	(32, 64, 96)	896
activation_1	(32, 64, 96)	0
convolution2d_2	(32, 62, 94)	9248
activation_2	(32, 62, 94)	0
maxpooling2d_1	(32, 31, 47)	0
dropout_1	(32, 31, 47)	0
convolution2d_3	(64, 31, 47)	18496
activation_3	(64, 31, 47)	0
convolution2d_4	(64, 29, 45)	36928
activation_4	(64, 29, 45)	0
maxpooling2d_2	(64, 14, 22)	0
dropout_2	(64, 14, 22)	0
flatten_1	(19712)	0
dense_1	(512)	10093056
activation_5	(512)	0
dropout_3	(512)	0
dense_2	(75)	38475
activation_6	(75)	0
	Total params	10197099

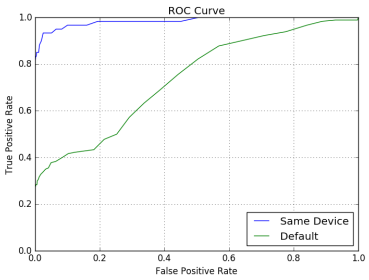

Fig. 10. Hybrid Model Receiver Operating Characteristics on MICHE-II

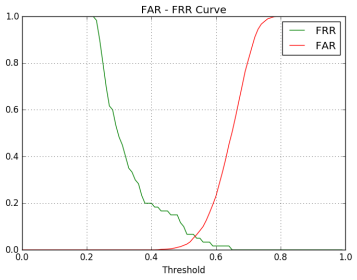

Fig. 11. Hybrid Model FAR-FRR in the SD_SE paradigm on MICHE-II

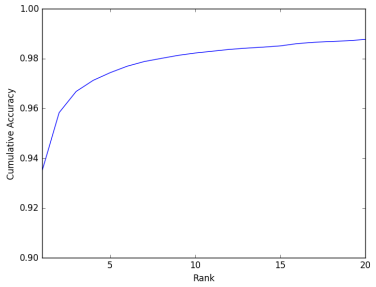

Fig. 12. Cumulative Match Curve for VISOB database

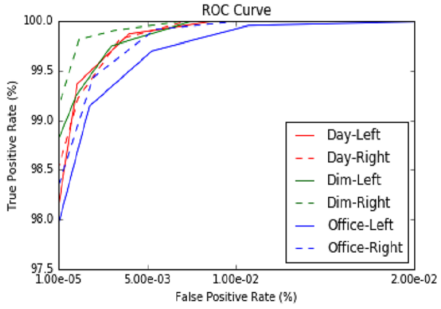

Fig. 13. ROC for Samsung on VISOB

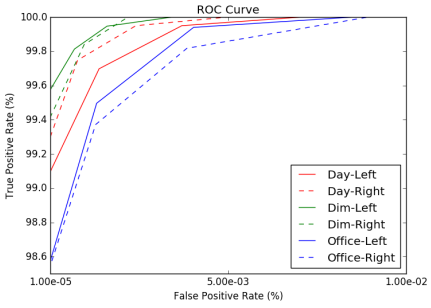

Fig. 14. ROC for Oppo on VISOB

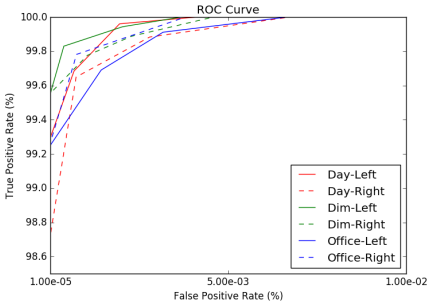

Fig. 15. ROC for I-Phone on VISOB

Table 3. Rank one accuracy (%) on VISOB database for different device and lighting condition

Phone	Condition	Left	Right
Samsung	Office	90.45	91.53
	Day	92.44	92.97
	Dim	93.12	93.61
IPhone	Office	93.54	93.89
	Day	95.98	94.82
	Dim	96.09	96.14
Oppo	Office	90.79	90.23
	Day	94.21	94.81
	Dim	96.31	96.15

6. Conclusion

In this paper, we proposed a hybrid convolution-based deep learning model, that combines a stacked unsupervised convolution-based model with a stacked supervised convolution-based model, and augments that with Root SIFT, for identifying an individual from a periocular image. This was obtained by training the underlying CNNs on a given set of periocular images as part of the learning phase, using transfer learning for using the features learned on external datasets for the case of the unsupervised convolutional network, and verifying a pair of images during the testing phase.

Our first unsupervised model, exploited *a-priori* knowledge to perform transfer learning, stemming from (a) a 128-dimensional facial feature vector exposed by OpenFace, and (b) a $1,024$ dimensional feature vector of the periocular region trained on VISOB database. It obtained similarity scores for each source-target pair using each of the two methods, used Root SIFT on the provided (MICHE-II) test data to obtain a dissimilarity score, and finally applied an average-based and a linear regression based score fusion technique to identify the best-matching source-target pair. The second model, on the other hand, used a 4-layer stacked convolution network followed by a 512-dimensional feature vector in a supervised learning paradigm, and used cosine similarity for testing purposes. VisobNet of the first model achieved a TPR of 99.5% at a FPR of 0.001% on the VISOB database, over all three smartphones and capture conditions. With respect to the MICHE-II test database, the first model produces a best-case AUROC of 0.956 and EER of 0.092, and the second produces a best-case AUROC of 0.981 and EER of 0.066, respectively. Both the ConvNets significantly outperform the baseline Root SIFT method, which yields a best-case performance of 0.453 and EER of 0.554. Model 2 outperforms the other models in a cross-device scenario, achieving an AUROC of 0.827 and EER of 0.271. The final model, a hybrid of the two convolution models with Root SIFT augmentation, is observed to deliver the best performance, under the constraint that the training and test data arise from the same device type, achieving an AUROC of 0.986 and EER of 0.053. The encouraging performance delivered by our approach, signify the potential of these models as candidates for deployment in real-life applications.

Table 4. Verification performance (TPR @ FPR = 10^{-3}) for different phones and capture conditions. Here BBSIF is Block BSIF, BHoG is to Block HoG, BSIF is Binary Statistical Image Features, HoG is Histogram of Gradients, LPQ is Local phase Quantization, DCA is Deeply Coupled Auto-encoders and DSF is Deep Sparse Filters

Feature	TPR(%)																	
	iPhone		Oppo		Samsung		iPhone		Oppo		Samsung		iPhone		Oppo		Samsung	
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
	Capture Condition : Day Light						Capture Condition : Dim Light						Capture Condition : Office Light					
BBSIF	45.77	42.69	46.22	49.40	47.63	48.56	40.05	35.93	26.62	51.77	44.44	48.56	29.47	30.71	26.62	23.30	24.45	30.27
BHoG	0.11	0.18	0.35	0.51	0.09	0.16	1.19	1.07	0.49	0.54	0.22	0.16	0.36	0.54	0.49	0.77	0.10	0.51
BSIF	60.11	61.52	54.40	53.45	54.39	62.93	43.74	48.61	28.00	56.76	54.82	62.93	42.14	44.45	28.00	30.68	34.29	39.64
HoG	0.04	0.03	0.19	0.18	0.06	0.13	0.41	0.53	0.32	0.30	0.36	0.13	0.28	0.33	0.32	0.37	0.24	0.34
LPQ	1.65	1.99	2.75	2.65	1.41	9.88	6.70	7.73	2.88	5.15	8.70	9.88	3.53	2.71	2.88	3.95	1.96	1.73
DCA	92.04	91.34	92.55	92.70	93.14	92.29	92.15	92.92	93.85	93.98	93.38	92.64	88.83	90.08	93.57	92.49	89.94	90.63
DSF	93.04	86.26	96.64	97.56	90.22	95.03	89.63	89.47	87.49	87.08	91.06	93.18	88.62	86.62	87.49	80.09	79.72	90.92
Our	99.75	99.67	99.53	99.77	99.27	99.14	99.87	99.76	99.85	99.84	99.24	99.71	99.55	99.79	99.27	99.18	98.62	98.89

Table 5. Equal Error Rate on MICHE-II

METHOD	CD_CE	CD_SE	SD_CE	SD_SE
Root SIFT	0.508	0.517	0.518	0.554
Visobnet	0.421	0.435	0.116	0.120
OpenFace	0.360	0.354	0.147	0.148
Model 1 LR	0.368	0.369	0.139	0.139
Model 1 EQ	0.409	0.417	0.106	0.092
Model 2	0.271	0.278	0.067	0.066
Hybrid	0.35	0.352	0.053	0.057

Table 6. AUROC on MICHE-II

METHOD	CD_CE	SD_SE	SD_CE	CD_SE
Root SIFT	0.500	0.453	0.486	0.486
Visobnet	0.637	0.924	0.928	0.623
OpenFace	0.619	0.924	0.922	0.694
Model 1 LR	0.691	0.956	0.956	0.688
Model 1 EQ	0.664	0.946	0.948	0.653
Model 2	0.827	0.981	0.984	0.815
Hybrid	0.751	0.985	0.986	0.736

References

[1] HYPR, Hypr, https://www.hypr.com/iot-security/.

[2] N. Maček, I. Franc, M. Bogdanoski, A. Mirković, Multimodal biometric authentication in iot: Single camera case study.

[3] U. Park, R. R. Jillela, A. Ross, A. K. Jain, Periocular biometrics in the visible spectrum, IEEE Transactions on Information Forensics and Security 6 (1) (2011) 96–106.

[4] C. Galdi, M. Nappi, J.-L. Dugelay, Multimodal authentication on smartphones: Combining iris and sensor recognition for a double check of user identity, Pattern Recognition Letters.

[5] U. Park, A. Ross, A. K. Jain, Periocular biometrics in the visible spectrum: A feasibility study, in: Biometrics: Theory, Applications, and Systems (BTAS'09), IEEE, 2009, pp. 1–6.

[6] D. L. Woodard, S. J. Pundlik, J. R. Lyle, P. E. Miller, Periocular region appearance cues for biometric identification, in: CVPR Workshops, IEEE, 2010, pp. 162–169.

[7] A. Rattani, R. Derakhshani, S. K. Saripalle, V. Gottemukkula, Icip 2016 competition on mobile ocular biometric recognition, in: ICIP, IEEE, 2016, pp. 320–324.

[8] R. Raghavendra, C. Busch, Learning deeply coupled autoencoders for smartphone based robust periocular verification, in: ICIP, IEEE, 2016, pp. 325–329.

[9] K. B. Raja, R. Raghavendra, C. Busch, Collaborative representation of deep sparse filtered features for robust verification of smartphone periocular images, in: ICIP, IEEE, 2016, pp. 330–334.

[10] K. Ahuja, A. Bose, S. Nagar, K. Dey, F. Barbhuiya, Isure: User authentication in mobile devices using ocular biometrics in visible spectrum, in: ICIP, IEEE, 2016, pp. 335–339.

[11] M. De Marsico, M. Nappi, D. Riccio, Noisy iris recognition integrated scheme, Pattern Recognition Letters 33 (8) (2012) 1006–1011.

[12] M. Haindl, M. Krupička, Unsupervised detection of non-iris occlusions, Pattern Recognition Letters 57 (2015) 60–65.

[13] M. De Marsico, M. Nappi, D. Riccio, H. Wechsler, Mobile iris challenge evaluation (miche)-i, biometric iris dataset and protocols, Pattern Recognition Letters 57 (2015) 17–23.

[14] K. Ahuja, R. Islam, F. Barbhuiya, K. Dey, A preliminary study of cnns for iris and periocular verification in the visible spectrum, in: ICPR, 2016.

[15] X. Cao, D. Wipf, F. Wen, G. Duan, J. Sun, A practical transfer learning algorithm for face verification, in: ICCV, 2013, pp. 3208–3215.

[16] B. Amos, B. Ludwiczuk, M. Satyanarayanan, Openface: A general-purpose face recognition library with mobile applications, Tech. rep., CMU-CS-16-118, CMU School of Computer Science (2016).

[17] R. Arandjelović, A. Zisserman, Three things everyone should know to improve object retrieval, in: CVPR, IEEE, 2012, pp. 2911–2918.

[18] I. Nigam, M. Vatsa, R. Singh, Ocular biometrics: A survey of modalities and fusion approaches, Information Fusion 26 (2015) 1–35.

[19] S. Crihalmeanu, A. Ross, Multispectral scleral patterns for ocular biometric recognition, Pattern Recognition Letters 33 (14) (2012) 1860–1869.

[20] M. De Marsico, C. Galdi, M. Nappi, D. Riccio, Firme: face and iris recognition for mobile engagement, Image and Vision Computing 32 (12) (2014) 1161–1172.

[21] M. Castrillon, M. De Marsico, M. Nappi, F. Narducci, H. Proena, Mobile iris challenge evaluation ii: Results from the icpr competition, in: ICPR, 2016.

[22] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: CVPR, 2014, pp. 1701–1708.

[23] F. Alonso-Fernandez, P. Tome-Gonzalez, V. Ruiz-Albacete, J. Ortega-Garcia, Iris recognition based on sift features, in: 2009 First IEEE International Conference on Biometrics, Identity and Security (BIdS), IEEE, 2009, pp. 1–8.

[24] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[25] M. Muja, D. G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration., VISAPP (1) 2 (331-340) (2009) 2.

[26] G. E. Dahl, T. N. Sainath, G. E. Hinton, Improving deep neural networks for lvcsr using rectified linear units and dropout, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 8609–8613.

[27] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Cognitive modeling 5 (3) (1988) 1.

[28] I. Sutskever, J. Martens, G. E. Dahl, G. E. Hinton, On the importance of initialization and momentum in deep learning., ICML (3) 28 (2013) 1139–1147.

[29] F. Chollet, Keras, https://github.com/fchollet/keras (2015).