Trends and Controversies

Hugo Proença
University of Beira Interior,
IT: Instituto de Telecomunicações

Mark Nixon
University of Southampton

Michele Nappi
Università di Salerno

Performing covert biometric recognition in surveillance environments has been regarded as a "grand" challenge, considering the adversity of the conditions where recognition should be carried out (e.g., poor resolution, bad lighting, off-pose and partially occluded data). This special issue compiles a group of approaches to this problem.

Progress in biometrics research has been concentrated on improving the robustness of recognition against poor quality data, consistent with less constrained data acquisition environments and protocols. Among the most obvious ambitions of this research topic is the development of automata able to work effectively in conditions that are currently confined to visual surveillance, so called "recognition-in-the-wild." In such conditions data is acquired covertly, from large distances, and has poor discriminability due to limited resolution, blur, and other degradation factors.

One interesting possibility to acquire data in visual surveillance scenarios is the use of PTZ (pan-tilt-zoom) devices. According to this concept, the QUIS-CAMPI surveillance system was recently introduced, enabling the automated acquisition of face imagery of subjects at-a-distance and on-the-move (up to 50 meters away). This dataset was the basis of the "ICB-RW: International Challenge on Biometric Recognition-in-the-Wild" competition, of which the primary goal was fostering the development of biometric recognition algorithms capable of working in surveillance scenarios.

The ICB-RW competition took place from September to December, 2015. There were a total of 19 registrations in the competition, most of these from academic/research institutions, also with a small number coming from private companies. A learning set from the QUIS-CAMPI database was initially released for all participants and, by the end of the contest, a disjoint subset was used in performance evaluation. Based on the obtained results, seven methods were selected, and their authors were invited to contribute to this department.

Ekenel et al. align the probe and gallery face images with respect to eye centers, considering only frontal images as gallery elements. A convolutional neural network (CNN) is used for face

representation purposes, with 1-nearest neighbor rule based on signal correlation being used for matching.

Grm and Struc generated an augmented version of the learning set by oversampling the training images via bounding box noise and horizontal flipping. The pre-trained Visual Geometry Group (VGG) face deep convolutional network was used as a feature extractor and a soft max classifier to discriminate between genuine and imposter pairwise comparisons.

Shi et al. used a feature set extracted from a deep convolutional network model trained on the CASIA-Webface database, and a similarity measure based on cosine distance. Ten models were learned independently from different facial parts, and subsequently fused. Also, multi-pose gallery data was synthesized to ease the matching phase.

Gutfeter and Pacut provided an information fusion approach that relies on the responses given by a set of convolutional neural networks that perform face recognition, each one specialized in handling samples from a specific 3D angle.

Brogan and Scheirer started by frontalizing both the gallery and probe data. Next, feature extraction was carried out based on a SLMSimple Neural Network with four bins created to represent different versions of the gallery samples. Finally, probe descriptors are matched with one of the four bins according to yaw angle of the head, and the resulting pairs of feature vectors feed a support vector machine that performs biometric recognition.

Gonzalez-Sosa et al. (Universidad Autónoma de Madrid, Spain) extracted Local Binary Patterns from nine facial regions of frontalized versions of the images. Next, illumination is compensated, and a fused distance score is determined by only considering the five best individual facial regions of each sample.

Finally, Riccio, Nappi, and de Maio started by locating a set of facial key points using an Active Shape Model. This step provided the information to remap (align) the face regions into 64 x 100 images of constant dimension. Next, local light adjustment techniques are used to compensate for the dynamic lighting conditions, with matching being carried out according to an optimized localized version of the spatial correlation index.

We hope that this collection of seven papers provides an overview of the current research in this extremely ambitious sub-field of biometric recognition research. We wish to thank all the people that enabled the publication of this special issue. First of all, we wish to thank Dr. Daniel Zeng, the editor-in-chief emeritus of this magazine, for accepting this idea with enthusiasm and for his support and motivation. Also, we would like to acknowledge the work carried out by João C. Neves, both in the management of the ICB-RW contest and in the performance evaluation of the submitted algorithms.

Acknowledgments

This work is supported by ‘’FCT – Fundação para a Ciência e Tecnologia” (Portugal), through the project “UID/EEA/50008/2013”.

DEEP REPRESENTATION AND SCORE NORMALIZATION FOR FACE RECOGNITION UNDER MISMATCHED CONDITIONS

Esam Ghaleb
Maastricht University

Gökhan Özbulak
Istanbul Technical University

Hua Gao
SensoMotoric Instruments (SMI)

Hazım Kemal Ekenel
Istanbul Technical University

Face recognition under unconstrained conditions is a challenging computer vision task. Identification under mismatched conditions, for example, due to difference of view angles, illumination conditions, and image quality between galley and probe images, as in the International Challenge on Biometric Recognition-in-the-Wild (ICB-RW) 2016, poses even further challenges.

In our work, to address this problem, we have employed facial image preprocessing, deep representation, and score normalization methods to develop a successful face recognition system. In the preprocessing step, we have aligned the gallery and probe face images with respect to automatically detected eye centers. We only used frontal faces as a gallery. For face representation, we have employed a state-of-the-art deep convolutional neural network model, namely the VGG-Face model. For classification, we have applied a nearest neighbor classifier with correlation distance as the distance metric. As the final step, we normalized the resulting similarity score matrix, which includes the scores of all face images in the probe set against all face images in the gallery set, with z-score normalization. The proposed system has achieved 69.8 percent Rank-1 and 85.3 percent Rank-5 accuracy on the test set, which were the highest accuracies obtained in the challenge.

Preprocessing

In the challenge dataset there are two subsets. These are gallery (watch list) and probe sets. The total number of subjects is 90. There are three face images for each subject in the gallery set and five images in the probe set. As gallery images, frontal, left, and right profile face images of the subjects, which are collected under control conditions, are provided. While, in the probe set, images of subjects collected from a surveillance camera are available.

In the proposed system, we only used frontal face images from the gallery set. We have also developed a multi-view based on another system,[1] however, due to low quality probe images, it is difficult to obtain good quality frontalized or profilized face images for matching. Moreover, we could not have achieved better results with this approach on the validation set.

Face alignment is based on eye-center positions. Even though we have tried to frontalize the probe images using the proposed method in Hassner et al.,[2] due to the aforementioned reason, this type of more advanced alignment did not provide a performance improvement on the validation set. Details of face alignment are given in the following subsections.

Facial Landmark Detection

For the given probe and gallery face image sets, 68 facial landmarks have been detected. The method in Kazemi and Sullivan[3] uses ensemble of cascade regression trees to estimate the facial landmark positions. Compared to other techniques, this method gives robust and accurate landmark positions in challenging conditions, such as varying illumination, pose, and low quality images, which are strongly present in the probe set of the ICB-RW competition.

Face Alignment

Face alignment is the process of registering faces with respect to facial landmarks, for instance, eyes, nose, mouth, and chin, to a canonical frame. This process fixes the landmarks' positions in aligned images, and it is carried out by a similarity transformation. In our work, we have used facial landmarks provided by the landmark detector in Kazemi and Sullivan[3] and performed 2-D similarity transformation that aligns faces based on eye center positions. After alignment, facial images were cropped and resized to a fixed resolution of 224 x 224 pixels.

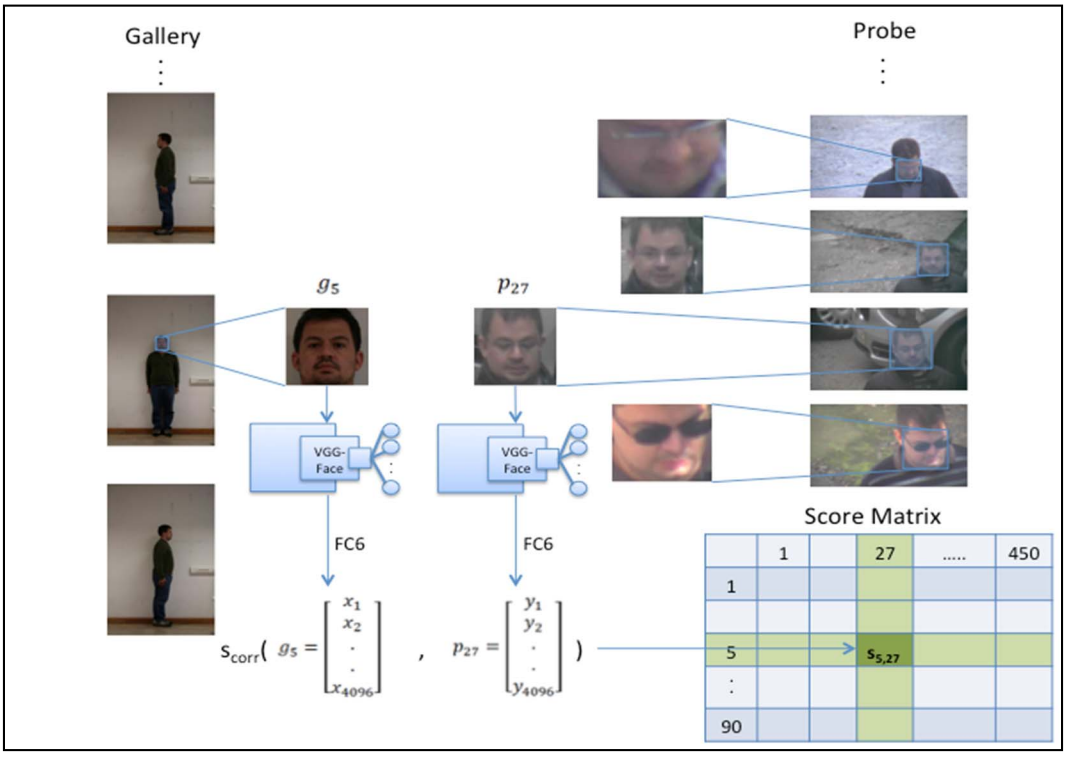

Figure 1. A general overview of our system.

Feature Extraction

Our face representation is based on VGG-Face model,[4] which is a 16-layer convolutional neural network (CNN) model trained with 2.6M facial images of 2,622 subjects. We used this model for feature extraction by employing the Fully Connected 6 (FC6) layer's output as the facial signature. This layer outputs a 4096 dimensional feature vector. The extracted feature vectors of the facial images in the gallery and probe sets are then compared using nearest neighbor classification with correlation distance as the distance metric. The overview of the system is illustrated in Figure 1.

Due to the limited amount of competition data, we did not perform fine-tuning to adapt the VGG-Face model,[4] which is trained mainly with the celebrity pictures collected from the web, to the competition's domain. However, the experiments on the validation set have shown that deep CNN-based representation still provides better performance compared to well-known approaches, such as Fisher vectors[5] (see Figure 2).

Post Processing

We have evaluated the similarity of every probe image against every gallery image and constructed a similarity score matrix with gallery images in the columns and probe images in the

rows. We first applied z-score normalization on each column of the score matrix, which represents the scores of gallery images for a given probe image, and then took the exponent of each matrix column in order to finalize the scores. The experimental results on the validation set have shown that such normalization has increased the Rank-1 accuracy from 66.8 percent to 71.7 percent.

Experiments

Challenge Task

Given one frontal, one left, and one right profile image for each of the 90 people in the gallery set with a total of 270 images, it is expected from the ICB-RW participants to evaluate the similarity scores of given probe images against the gallery images (watch list). There are five probe images for each person with a total of 450 images provided for validation and test, respectively. As a result, the system provides an MxN dimensional score matrix, where M is the number of face images in the gallery set and N is the number of face images in the probe set. Since, in the proposed system, only frontal face images of subjects were used from the gallery set, M corresponds to the number of subjects, 90, in the database.

In the score matrix, each column represents the distance values between a subject's probe image and all the images in the gallery set. A score value at the (i,j)-th position in the score matrix corresponds to the distance between the i-th gallery image and the j-th probe image.

Evaluation

As required in the competition, the algorithm's performance is measured by the Rank-1, Rank-5 accuracies and Area Under Curve (AUC) of the Cumulative Match Score Curve (CMC). For each probe image, a Rank-K list is calculated by ranking the K most similar subjects in the watch list. The CMC is obtained by calculating the percentage of correct identification for all probe images with all different Rank-K list sizes.

The CMC plots obtained by all the tested features on the validation set can be examined in Figure 2. On the validation set, with the VGG-Face features, 66.8 percent Rank-1, 83.5 percent Rank-5 accuracies and an AUC of 95.3 percent is obtained. With the score normalization, Rank-1 and Rank-5 accuracies increase to 71.7 percent and 86.5 percent, respectively. AUC has become 96.2 percent. The AUC performance of Fisher vectors is 85.6 percent and increases to 86.5 percent when normalization is applied.

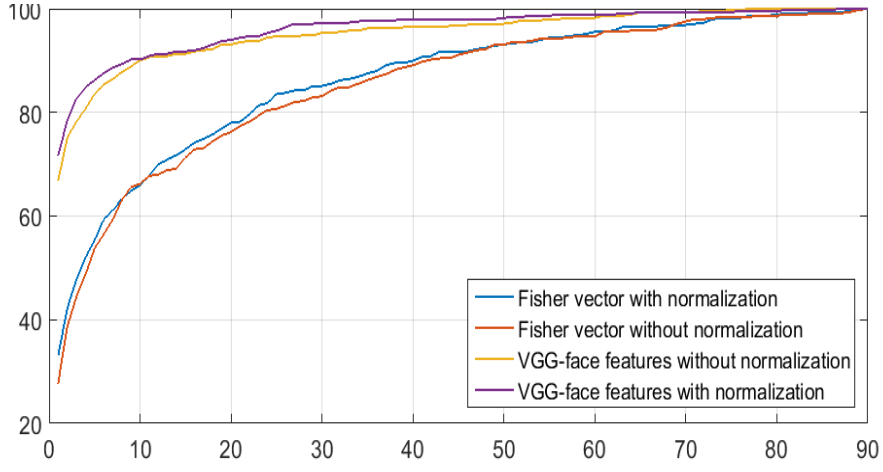

Figure 2. Cumulative Match Score Curve (CMC) on the validation set.

The AUC obtained on the test set is 95.4 percent. A Rank-1 accuracy of 69.8 percent and a Rank-5 accuracy of 85.3 percent have been achieved. These values are the highest accuracies obtained in the challenge. Compared to the Rank-1 scores of the second and third best systems in the competition, our system provides 7.8 percent and 12.2 percent absolute performance improvement, respectively. The results validate the difficulty of performing face recognition under mismatched conditions, which indicates that further research is required to improve performance.

Acknowledgments

This work was supported by TUBITAK project number 113E067 and by a Marie Curie FP7 Integration Grant within the 7th EU Framework Programme.

DEEP FACE RECOGNITION FOR SURVEILLANCE APPLICATIONS

Klemen Grm
University of Ljubljana

Vitomir Struc
University of Ljubljana

Automated person recognition from surveillance quality footage is an open research problem with many potential application areas. In this paper, we aim at addressing this problem by presenting a face recognition approach tailored towards surveillance applications. The presented approach is based on domain-adapted convolutional neural networks and ranked second in the International Challenge on Biometric Recognition-in-the-Wild (ICB-RW) 2016. We evaluate the performance of the presented approach on part of the QUIS-CAMPI dataset and compare it against several existing face recognition techniques and one state-of-the-art commercial system. We find that the domain-adapted convolutional network outperforms all other assessed techniques, but is still inferior to human performance.

The demand for surveillance systems is growing rapidly. To be useful, such systems require active human supervision and screening of all recorded surveillance footage, which is a demanding and time-consuming task considering the number of security cameras commonly installed at the surveilled areas. Clearly there is a need to devise automated approaches capable of autonomously recognizing people from security videos without human intervention. Unfortunately, the quality and variability of the security footage makes it difficult to develop automated solutions capable of matching human performance. To address this problem, we present in this paper a face recognition approach based on domain-adapted convolutional neural networks. The presented approach exploits the so-called VGG convolutional network trained on a large dataset of facial images and uses the pre-trained VGG network to process the security footage and extract high-level facial representations. A softmax classifier is then trained on top of the deep network using facial images captured by a security camera. Here, the classifier acts as a domain-adaption layer which exploits the facial representations produced by the network to conduct identity interference in the target domain (i.e., on the security footage). In the remainder of the paper we describe the domain-adapted convolutional network used for our ICB-RW submission and present experimental results on the QUIS-CAMPI[6] dataset. We describe comparative experiments with various face recognition systems and also compare the performance of the presented approach with human performance on the same data.

Deep Learning for Surveillance Applications

Deep Learning and Convolutional Neural Networks

In recent years, deep learning has attracted significant attention in various application domains, such as natural language processing, computer vision, or signal processing. Deep models have

shown state-of-the-art performances for different research problems by learning high-level feature representations from raw input data through a hierarchy of model layers. For computer vision problems, the predominant deep models are convolutional neural networks (CNNs), which consist of cascaded stacks of convolutional filters. The networks as a whole are parameterized by the weights of the individual filters $\theta = \{W\}$ that are learned during training. At each layer, the output of the previous layer is processed via convolutional filtering, and the output is subjected to a non-linear activation function. For the n-th layer of an N-layer network, this can be formalized as follows:

$$y_n = f_{\theta_n}(y_{n-1}) = \sigma(y_{n-1} * W_n) \qquad (1)$$

where y_n and y_{n-1} ($1 \leq n \leq N$) represent the outputs of n-th and ($n-1$)-th layer, respectively; σ denotes a non-linear activation function; $*$stands for the convolutional filtering; the set of open parameters of the n-th layer are the filter weights, i.e., $\theta_n = \{W_n\}$; and the input to the first layer ($n = 1$) are the raw (unprocessed) images. An N-layer deep CNN is then described as:

$$y = \left(f_{\theta_N} \circ f_{\theta_{N-1}} \circ \ldots \circ f_{\theta_1}\right)(x) \qquad (2)$$

where x and y are inputs and outputs of the network, respectively, and $\circ$ stands for the function-composition operator. To reduce the computational requirements and the size of the parameter space of the CNNs, the convolutional layers are commonly interspersed with dimensionality-reducing layers, such as max-pooling, average pooling, or strided convolutional layers, which effectively implement different subsampling strategies. By training convolutional networks via gradient descent, the image representation is learned directly from the input data in an end-to-end manner, as opposed to classical computer vision approaches where the image descriptors are typically hand-crafted before being fed to some classifier.

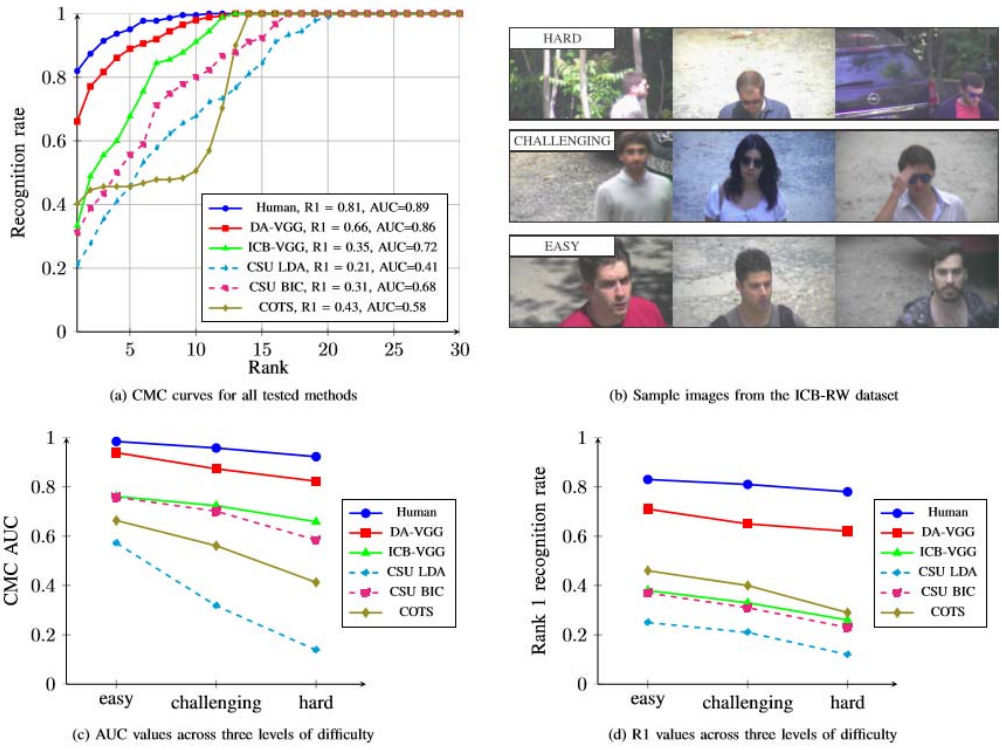

Figure 1. Experimental results of the evaluation. The images show: (a) the CMC curves for the comparative assessment; (b) sample images from the ICB-RW dataset (manually) partitioned into three subsets according to the level of difficulty the images pose for the recognition process; (c) a

comparison of AUC values across the three difficulty levels for all assessed methods; and (d) a comparison of Rank 1 recognition rates across the three difficulty levels for all assessed methods.

The VGG architecture

The VGG network architecture, introduced for face recognition in Parkhi et al.,[4] represents a 16-layer CNN that falls into the class of so-called very deep convolutional networks. The VGG network achieves competitive performance due to some key differences over earlier network architecture, i.e.:

- Small filters: All convolutional filters are of size 3 x 3 pixels, as opposed to earlier CNNs which used much larger filter sizes. By using multiple 3 x 3 convolutions in a sequence, a similar effect is achieved as with larger filters (receptive fields), but with a less extensive parameter space.
- No strides: Previous CNN implementations used large filters combined with strides of more than 1 (commonly: 4) to subsample the input image. This adversely affects performance and is not required with the VGG architecture.
- Constant representation size: Every sub-sampling step by a factor of 4 (max-pooling over a 2 x 2 neighborhood) is followed by a 2-fold increase in the number of convolutional filters in the following layers. This process results in a constant representation size of all layer outputs (in terms of memory requirements) and improves the computational performance of the CNN.

The VGG network for surveillance applications

Training a competitive VGG network for face recognition in surveillance scenarios requires large amounts of training data and significant computing resources. The original VGG network, for example, was trained with 2.6 x 10^6 facial images over several weeks on a computer equipped with 4 high-performance GPUs.[4] To make the VGG network applicable to surveillance scenarios, we resort to domain adaptation techniques and apply them to the pre-trained VGG (face) convolutional network from Parkhi et al.[4] We perform net surgery on the pre-trained VGG network and use the existing configuration for representation calculation. Specifically, we use the output of the final fully-connected layer as the representation of the input images. On top of the network (i.e., after the fully-connected layer) we train a probabilistic multi-class softmax classifier using the development set of the ICB-RW data.

Assume a set of training vectors $y = \{y_i\}_{i=1:L}$ belonging to M distinct classes. A softmax classifier computes a vector of posterior probabilities $p \in \mathbb{R}^{M\times 1}$ for all target classes through the softmax transformation of a linear function of y, i.e.:

$$\mathrm{p} = \frac{e^{\mathrm{W}^T \mathrm{y}+\mathrm{b}}}{\sum_{i=0}^{M} e^{\mathrm{W}_i^T \mathrm{y}+\mathrm{b}_i}}$$

where the image representation $y \in \mathbb{R}^{K\times 1}$ is generated by the pre-trained VGG network, and the matrix $\mathrm{W} = \left[\mathrm{w}_1^T, \mathrm{w}_2^T, \ldots, \mathrm{w}_M^T\right]^T \in \mathbb{R}^{K\times M}$ and the vector $\mathrm{b} = \left[b_1, b_2, \ldots, b_M\right]^T \in \mathbb{R}^{M\times 1}$ are learned parameters of the classifier. The classifier is trained via mini-batch error backpropagation with stochastic gradient descent using the categorical cross-entropy between the current output probability distribution and the desired probability distribution as the objective function. A given input vector y is classified into the class with the highest posterior probability. With the presented approach, the pre-trained VGG network is treated as a feature extractor and the softmax classifier as the domain-adaptation layer that maps the computed image representation into the target application domain. We refer to this approach as the domain-adapted VGG network (DA-VGG) in the remainder of the paper.

Experiments and Results

We assess the suitability of the domain-adapted VGG network for surveillance scenarios on part of the QUIS-CAMPI[6] dataset used for the ICB-RW 2016 competition. The data contains gallery and probe images of 90 distinct subjects (See Figure 1b). The high resolution gallery images consist of one frontal and two profile images of each subject captured under frontal pose and uniform illumination in studio-like conditions. The probes are of lower quality and comprise 10 images captured by a security camera. Our goal is to automatically determine the identity of the subjects in the surveillance footage (i.e., the probes) given the high-resolution galleries. For the experimental evaluation, we follow the ICB-RW protocol and split the gallery and probe images into a development set, used for training, and an (hold-out) evaluation set, used for performance reporting. The former contains all gallery images and half of the probes, which the latter comprises the same galleries and the other half of the probe images. We conduct 450 identification experiments (each involving 270 probe-to-gallery comparisons) for each experimental run. We report performance in terms of Cumulative Match Score Curves (CMCs), the rank-R1 (R1) recognition rate and the area under the CMC curves (AUC). Prior to the experiments, we crop facial regions from the gallery and probe images using the bounding boxes that ship with the data and rescale the cropped regions to a size of 224 x 224 pixels.

We provide competitive results for a number of competing methods, i.e:

- CSU baseline recognition systems based on Linear Discriminant Analysis (CSU LDA) and the Bayesian intrapersonal/extrapersonal classifier (CSU BIC)[7]
- A deep convolutional neural network based on the VGG architecture was trained from scratch in the learning set of ICB-RW data (ICB-VGG) and
- A state-of-the-art commercial off-the-shelf (COTS) face recognition system.

Additionally, a trained researcher manually assigned a similarity score between 1 (surely different people) and 5 (surely the same person) to each probe-to-gallery comparison to provide insight into the capabilities of human annotators on the data. The scoring methodology followed the approach presented in Phillips and O'Toole,[8] and the results generated based on this scoring are denoted with "Human" in Figure 1. The CMC plots of the experiments are presented in Figure 1(a). The DA-VGG network outperforms the CSU baselines with a margin of over 30 percent in terms of the rank-1 recognition rate. The domain adapted network (DA-VGG) also results in better performance than the ICB-VGG network trained from scratch, suggesting that large amounts of training data (albeit outside the problem domain) are a must for the training of competitive deep models. The COTS system results in a rank-1 recognition rate of 43 percent, which is below the 66 percent assured by the DA-VGG network. However, facial detection is an integral part of the COTS-system, so the reported performance also includes potential errors at the face detection stage, which is not the case for other methods. Among all tested approaches, the DA-VGG performance is the closest to human performance, though the performance gap is still around 15 percent on this dataset at rank-1 in favor of humans. This observation is in line with previous work,[8] which also suggests that for difficult conditions automatic systems are still inferior to humans. To further break down these results, a human annotator partitioned all the probe images into three subsets (i.e., easy, challenging, and hard) according to the perceived level of difficulty of the images for recognition illustrated in Figure 1(b). The AUC values and rank-1 recognition rates across the three levels are shown in Figures 1(c) and 1(d) for all assessed methods. The human performance is the most consistent, while all other methods deteriorate in performance when moving to more difficult conditions. In terms of AUC, human and DA-VGG performance are reasonably close on the "easy" images," while the performance gap is bigger for the "hard" images.

Conclusions

We have presented our work related to the ICB-RW evaluation. Our experimental results suggest that, despite the lack of large-scale datasets of surveillance footage suitable for training deep face recognition models, adaptation techniques can be exploited to develop models with reasonable performance. Nevertheless, automated face recognition for surveillance applications remains a

challenging problem, and human performance still remains superior for difficult conditions. Given the potential benefits of fully-automated surveillance systems, further research in this area is warranted.

DEEP NETWORK ENSEMBLE FOR SURVEILLANCE FACE RECOGNITION

Hailin Shi
Xiangyu Zhu
Shengcai Liao
Zhen Lei
Stan Z. Li
Chinese Academy of Sciences

Surveillance face recognition is important for watch-list based applications. However, face recognition in surveillance scenarios is difficult due to various challenges. It is a kind of "in the wild" scenario, but generally more difficult than face recognition with Internet images (e.g., labeled faces in the wild).

In this paper, we introduce a deep network ensemble method for surveillance face recognition. We adopt a deep model of convolutional neural networks (CNN) as the feature descriptor and make an ensemble of ten such models learned from different facial parts. We also propose a multi-pose synthesis method to expand gallery images for better matching.

Model

Our CNN model includes 9 convolutional layers and 4 pooling layers, without any fully-connected layer to keep the model in a light-weight style. The architecture details are given in Table 1, determined similarly as in Yi et al.[9]

Table 1. The architecture of the CNN.

Name	Type	Filter / Stride	Output
Conv11	Convolution	3×3 / 1	55×55×32
Conv12	Convolution	3×3 / 1	55×55×64
Conv13	Convolution	3×3 / 1	55×55×128
Pool1	Max pooling	2×2 / 2	28×28×128
Conv21	Convolution	3×3 / 1	28×28×96
Conv22	Convolution	3×3 / 1	28×28×192
Pool2	Max pooling	2×2 / 2	14×14×192
Conv31	Convolution	3×3 / 1	14×14×128
Conv32	Convolution	3×3 / 1	14×14×256
Pool3	Max pooling	2×2 / 2	7×7×256
Conv41	Convolution	3×3 / 1	7×7×160
Conv42	Convolution	3×3 / 1	7×7×320
Pool4	Avg pooling	7×7 / 1	1×1×320
Norm4	L2 normalization		320

Training

Our CNN model is trained on the CASIA-Webface database,[9] which contains 10,575 subjects and 494,414 face images. We sample a validation set with 1,000 subjects from the database, and use the remaining subjects for training. All face images are aligned according to the provided landmarks, and then horizontally mirrored for data augmentation.

The proposed CNN model is trained in a multi-task fashion.[9] The training objective is a weighted sum of the softmax and contrastive losses[10] employed after the Norm4 layer of the CNN. This incorporates both the advantages of face identification and verification tasks for joint learning.

The softmax part is conducted by an N-way classification of the training identities. The CNN features and the identity labels are involved to calculate the log-likelihood loss. This part is in charge of learning discriminability of the CNN via the identification task.

Meanwhile, the contrastive loss is responsible for learning generalization ability of the CNN via the verification task. The contrastive loss[10] is computed by an L2 distance D^2 if the pair of face features is positive (the same subject), and by a hinge loss otherwise (Equation 1). Through the contrastive loss learning, the CNN tends to reduce the distance of the positive pairs, and push the negative pairs away.

$$\text{contrastive loss} = \begin{cases} \frac{1}{2}D^2 & \text{, Positive pair} \\ \frac{1}{2}\max\left(0, 1-D\right)^2 & \text{, Negative pair} \end{cases}$$

The training is conducted in the manner of mini-batch optimization. The CNN is optimized via the standard stochastic gradient descent method with back-propagation. According to the training steps in Yi et al.,[9] we set the softmax loss a large weight in the beginning of training and reduce it iteratively because the softmax loss converges much faster than the contrastive loss.

Multi-pose Synthesis of Gallery Images

In the evaluation phase, we adopt a newly proposed face synthesis approach[11] to enlarge the gallery set so as to improve the robustness against pose variations. Specifically, given three images (frontal, left and right profiles) of each subject in the gallery set, firstly, we detect 68 facial landmarks of each image by Yan et al.[12] and fit a 3DMM model to the three images, which are constrained to have different poses but share the same 3D shape. Secondly, for each image, we uniformly mark some background anchors around the face region, estimate the depth with their nearest 3D points, and turn the whole face image into a 3D object. Next, the three 3D faces are merged by the dense correspondence in face region, leading to a refined 3D description of the gallery face. Finally, we rotate the 3D gallery face in 54 poses (9 yaw x 6 pitch) and render them into 2D images to expand the gallery set, as shown in Figure 1.

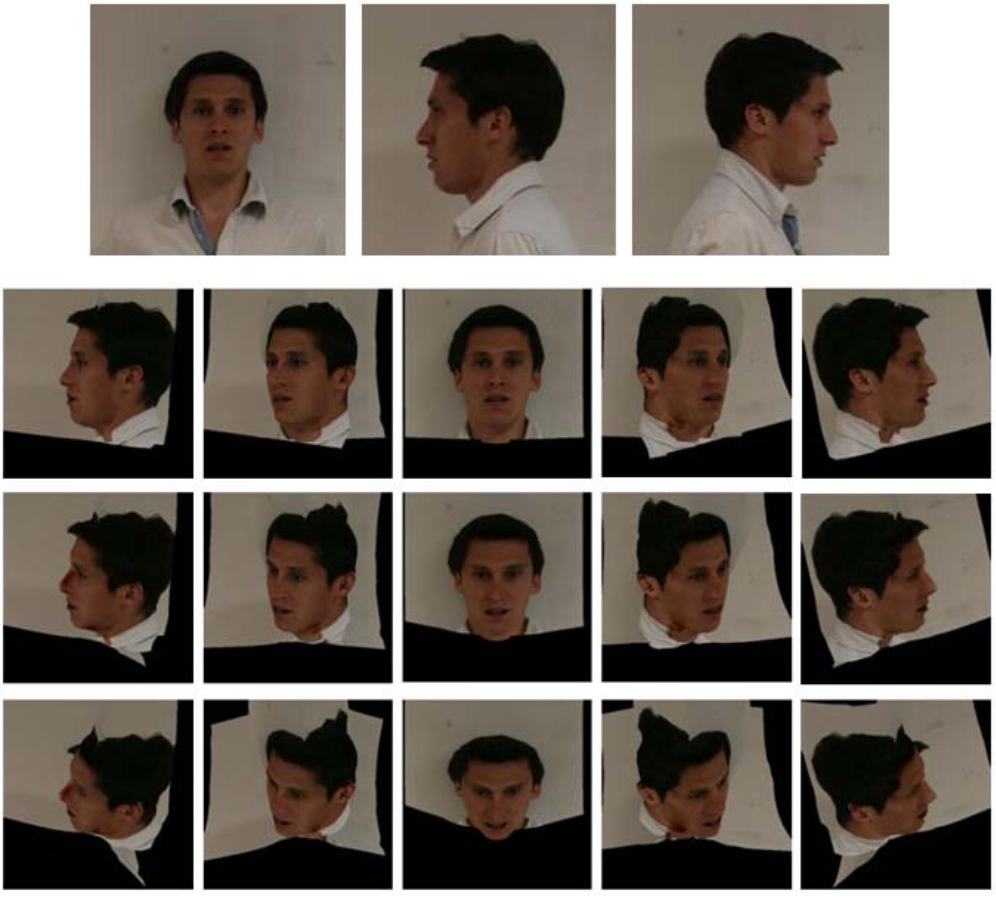

Figure 1. Original gallery images (top) and some examples of synthesized gallery images of different poses (bottom).

Evaluation

The evaluation process is described as follows. First, the face area and landmarks are detected by Yan et al.[12] from the raw gallery and probe images.

Then, all the gallery (including synthesized ones) and probe images are aligned and cropped into 55 x 55 RGB images according to the detected landmarks. These images are sent to the CNN model for feature extraction, resulting in a 320-d feature vector for each image.

Next, a cosine similarity matching between the feature vectors is performed for the matching score. In particular, the scores between the probe and the synthesized gallery images are averaged as the matching score between the probe and the gallery subjects. Moreover, ten of such CNN models are trained, each focusing on different facial parts. Finally, scores of the ten models are averaged to get an ensemble and yield the final result.

The proposed method is evaluated on the International Challenge on Biometric Recognition-in-the-Wild (ICB-RW). This benchmark dataset contains 90 subjects, with three images (frontal, left, and right profiles) per subject enrolled in the gallery set, five images per subject in the probe set for training, and another five images per subject in the probe set for performance evaluation.

Following the evaluation protocol, our method achieves 57.6 percent rank-1 accuracy and 0.921 AUC of the CMC curve, leading to the 3rd place in the ICB-RW competition (Table 2). Nevertheless, it is worth noting that the parameter size of our model is 17.2 MB, which is much lighter than the VGG model (37.8 MB) used by the top two methods in Table 2.

Table 2. Performance on the ICB-RW.

Method	Rank-1 IR (%)	Rank-5 IR (%)	AUC (CMC curve)
H. Ekenel, G. Ozbulak, E.Ghaleb	69.8	85.3	0.954
K. Grm, S. Dobrisek, V. Struc	62.0	78.7	0.952
H. Shi, X. Zhu, S. Liao, Z. Lei, S. Li	57.6	75.8	0.921
W. Gutfeter	42.9	64.4	0.918
J. Brogan	11.6	30.4	0.755

It appears that the face recognition performance in surveillance is promising here, thanks to the powerful CNN models. However, keep in mind that this is a small dataset, therefore a large benchmark on surveillance face recognition is still an urgent need.

DEEP NEURAL NETWORK ENSEMBLES DEDICATED TO DIFFERENT HEAD POSES FOR FACE IDENTIFICATION

Weronika Gutfeter
Research and Academic Computer Network (NASK)

Andrzej Pacut
Warsaw University of Technology

A solution is proposed to the problem defined in the International Challenge on Biometric Recognition-in-the-Wild (ICB-RW 2016). Faces from images taken under uncontrolled conditions are to be identified in a watch-list created of good quality facial images of various head poses. The proposed method is based on deep neural network ensembles. Each network ensemble is trained to classify faces seen from specific directions, namely, frontal images, left, and right profiles. This approach reflects the watch list structure, which contains these three types of head poses. The ensemble networks' results are then combined in various ways to obtain the final classifier. The results are compared to a single network trained with all pooled facial images.

The Objective of Challenge and Specification of Data

The algorithm described in this paper was designed to meet the objectives of the ICB-RW 2016 challenge. The task of this competition was to identify the people in the CCTV surveillance system assuming that one is provided a watch list which consists of good quality images. The training data used in the competition was a part of the QUIS-CAMPI dataset. It contains both the watch-list images of 90 subjects (called there the gallery subset) and the probe images which are static frames obtained in a surveillance system. In the gallery subset, three images with different head positions (frontal, left-side, and right-side) were provided for each subject. The gallery images were acquired under controlled lighting conditions, and aside from faces, they show full silhouettes of the subjects. The probe images subset contains five images of each subject captured outdoors. The subjects rarely look into the camera that was localized over their heads, and faces are often occluded. Challenge participants received the gallery subjects' detection results, namely, the positions of detected faces. The results of the ICB-RW Challenge were evaluated

using the area under curve (AUC) calculated for the cumulative match score (CMC) as an identification rate. The participants had no access to an additional five probe images that were kept for evaluation purposes.

Deep Learning Methods in Face Recognition and Image Classification

Methods based on deep convolutional neural networks became winning approaches in various competitions in the field of image classification and face recognition. Training multilayer networks is a very time- and resource-consuming process. The concept behind convolutional networks lies in the reduction of the connections between layers – each convolutional layer is a set of small-neighborhood filters moving through the layer and having shared parameters that are not dependent on filter positions. Typically, the convolution layers are separated by the pooling layers which aggregate the results of convolutions. Usually, the last layers are fully connected, and the output layer expresses the class probabilities. It is believed that the features built in the consecutive layers express the increasingly complex image properties relevant to a given recognition problem.

Architecture of Network

Three networks were built, each trained on a subset of watchlist images: frontal, left-sided, and right-sided. The results were merged in different ways to obtain final classifiers. Several merging methods were compared along with a pooled data classifier, trained on all (frontal, left-sided, and right-sided) images. The networks were based on certain recommendations given in some other papers describing image recognition methods based on deep neural networks Very effective solutions using convolutional neural networks were introduced in the paper; they are referred to here as VGG (Visual Geometry Group) networks. The VGG networks are built of a set of layers of convolutional filters of different sizes. The network architecture is similar in structure to CNN-S, which is the slowest but also the most accurate version of the proposed networks.

Some layers were reduced to better meet the challenge requirements and also to overcome certain hardware limitations of the laboratory equipment. The network processes 3-channel RGB image patches of width and height of 128 pixels; hence, it accepts the inputs of size 3 x 128 x 128. The network is comprised of 5 convolutional layers, one with filters of size 7 x 7 (number of filters was set to 96), one of size 5 x 5 (256 filters), and the last three of sizes 3 x 3 (512 filters in each layer). The convolutional layers were separated with pooling layers consisting of 2 x 2 or 3 x 3 not-intercepting pooling regions. The number and sizes of the inner layers remain the same as in the original VGG architecture. The last 2 layers were fully connected and had their sizes reduced from the original 4096 units to 1024 units. At the output, the probabilities of 90 subjects were calculated. The network was implemented using common deep learning libraries in python programming language: lasagna and low-level theano. They have the ability to transfer some CPU computations to GPU processors to accelerate the learning process. A typical training took 11 hours for each network.

Data Preprocessing and Network Training

Data preprocessing has a strong influence on the convergence of the training process. First, the data was cropped using the detection coordinates provided together with the images. The cropped images were then processed by histogram normalization using its luminance component in YCbCr color space (only the luminance channel was normalized). To increase the number and variability of the training set, the original gallery dataset was augmented by small random transitions, small rotations, and pixel value dithering, giving 36 images for each gallery image. Then the images were scaled to size 128 x 128 (with the ratio preservation). Afterwards, the pixel values were demeaned and scaled to a common range. The augmented gallery set contained 9,720 images, 3,240 images for each head position. Five probe images were available for each person.

Four probe images and all relevant face images for network training were used, leaving the remaining probe image for validation. All frontal, left-side, and right-side, images were classified as relevant, depending on the type of training. Each training was repeated 7 times, using different folds of the probes data. Training batches contained 45 images, due to limitation of the memory of the graphic card. The stochastic gradient descent with the momentum was chosen for learning. The hyper-parameters were changed linearly through the learning epochs, with the momentum increasing from 0.9 to 0.999 and with the learning speed decreasing from 10^{-2} to 10^{-4}. The validation error was defined by the categorical cross-entropy

$$E = -\sum_{ij} t_{ij} \log p_{ij}$$

Where p_{ij} is the predicted probability that i-th image belongs to j-th class and t_{ij} is defined by the target distribution.

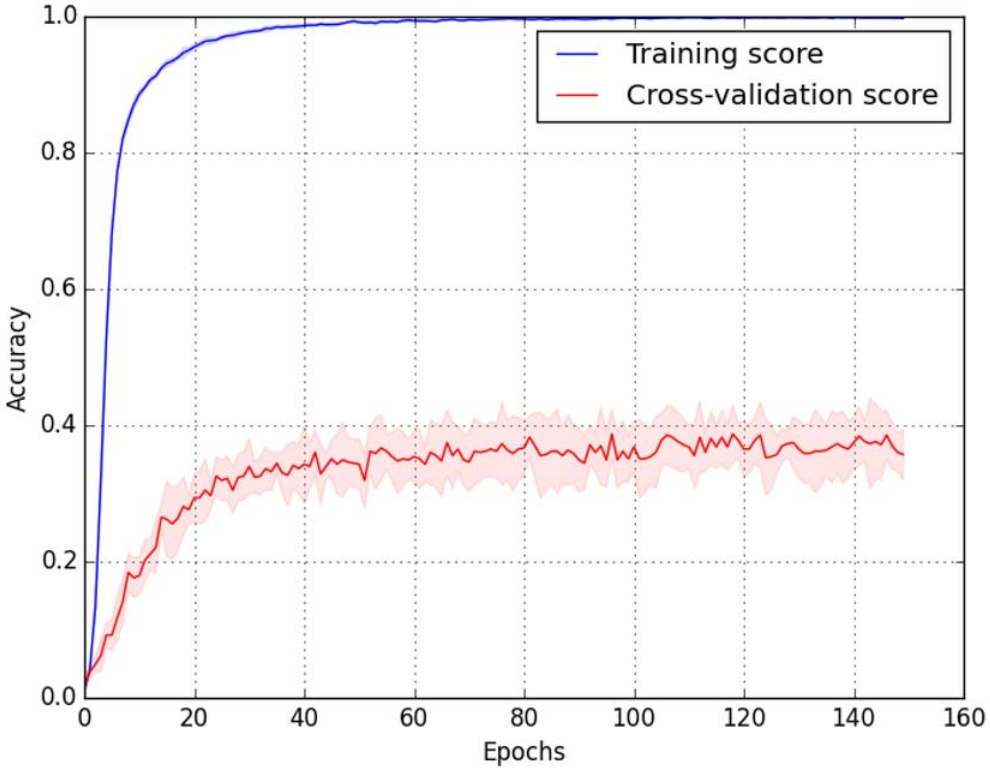

Figure 1. The training process: Accuracy for the estimation and the validation sets. The shaded area shows the standard deviation limits.

After the initial tests, the maximal number of epochs was set to 150. This number was determined experimentally.

Results and Discussion

Each single neural network (for frontal, left-side, and right-side faces) produces a vector of probabilities of the size equal to the number of classes. They can be aggregated in various ways. Some simple methods of merging the predictions from the trained ensembles were tested on the networks. The maximal value, the minimal value, and the average were compared in an experiment. A network on all pooled data was also trained. The resulting scores are shown in Table 1. The table presents the average results for all 7 folds completed during training. As the challenge

was evaluated by the area under curve (AUC), it was decided to use the average of network predictions as the final merging method.

The difference between the cumulative curves of identification between single network and the average result for ensemble can be seen in Figure 2. Note that while the pooled data results are similar to the individual (F, L, or P) results, they were obtained with a three-times larger dataset. After merging the results (Figure 2), the CMC curves for the ensemble average are larger than the individual results by about 0.1.

Table 1. The scores for F, L, R, and pooled data networks, together with the scores for several ways of predictions merging.

Ensemble	Rank-1	Rank-3	Rank-12	AUC
frontal (F)	0.344	0.508	0.694	0.8362
left (L)	0.303	0.465	0.678	0.8208
right (R)	0.322	0.467	0.670	0.8158
pooled	0.321	0.498	0.671	0.8164
average (F,L,R)	0.383	0.535	0.775	0.8939
median (F,L,R)	0.389	0.563	0.751	0.8561
max (F,L,R)	0.363	0.532	0.757	0.8911
min (F,L,R)	0.389	0.492	0.598	0.7645

This suggests that using the ensemble approach by a proper "segregation" of data leads to better classification results.

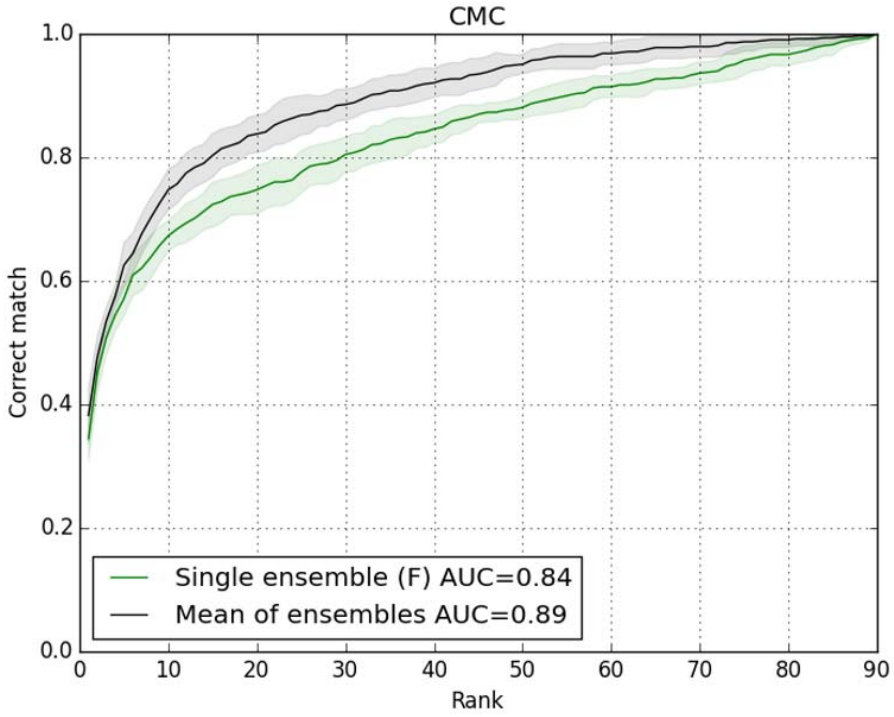

Figure 2. Cumulative match curves for a single network (F) and for the average of ensembles.

Conclusions

It was shown that the convolutional neural networks can be robust classifiers in the task of human face identification in challenging conditions. Without additional knowledge and pre-training on external databases, the system was prepared to recognize faces of the subjects in distinct poses. The classification given by the ensemble approach (with the use of averaging the individual results) is better than that given by pooled data. The presented approach leads to one of the best in the ICB-RW 2016 challenge.

FACIAL FRONTALIZATION AND SMART MATCHING VIA POSE

Joel Brogan
Walter J. Scheirer
University of Notre Dame

In this work, we introduce a face recognition method based on the idea of improving the performance of a deep convolutional neural network by frontalizing and accurately aligning extreme out-of-pose images as a pre-processing step before feature extraction. Highly accurate facial alignment and pose normalization provide spatially coherent feature patches for faces of different shape and pose.

Methodology

This method has three main components: (1) Facial pose correction and binning, (2) feature extraction via a biologically-inspired convolutional neural network, and (3) face matching using an SVM trained on comparison vectors.

Facial pose correction and binning

For facial pose correction and alignment, a modified version of Hassner's method[2] was used. Hassner's frontalization method normalizes pose by calculating an extrinsic camera calibration. Using a 68-point facial landmarker from Zhu and Ramanan,[13] the detected points on the input face are compared to a set of template points on a generic 3D face template to calculate a 2D to 3D transform. This transform is used to back-project face pixels from the image onto a frontalized generic 3D face model. This method helps accurately align features on out-of-pose faces for better matching performance. To address both extreme and mild out-of-plane facial rotation, four different watch-list "bins" of frontalized faces are built from the watch-list set. Bin 1 consists of versions of the face frontalized with no symmetry induced. Bin 2 consists of versions of the face replacing areas that are self-occluded from heavier yaw angles using a "soft symmetry" replacement mask. Bin 3 consists of versions of the face with complete symmetry mirrored from the left side to the right. Bin 4 is the same as Bin 3, using the opposite side of the face for symmetry. This 4-bin database system induces all possible variations of the frontalized watch-list set that account for different types of poses. In this way, it is less likely for a probe face to be at a pose unaccounted for within the system. When an input probe is assigned to the correct database bin, it will match against face images frontalized under similar conditions to improve match performance.

Feature extraction

Once the watch-list pose database is built, a biologically-inspired artificial neural network is used to extract a rich feature vector from the face.[14] This network is a three-layer convolutional neural network trained by randomly instantiating the weights for a large set of candidate models, with selection of the best models made via a high-throughput screening approach. As a starting point, and inspired by previous neuronal modeling work in computational neuroscience, the

model considers the constituent operations of cortical processing in a single layer as a set of simple computational elements, including (i) a filtering operation, implementing template matching; (ii) a simple nonlinearity, e.g., a threshold; (iii) a local pooling/aggregation operation, such as softmax; and (iv) a local competitive normalization. Each of these operations is actually a large family of possible operations, specified by a set of parameters controlling fan-in and fan-out, threshold values, pooling exponents, the spatial extent over which the operations perform, and the size/shape/content of the templates that are matched. A simulated cortical unit is then modeled as a specific choice of these elements, e.g.:

$$Simulated\ unit\ output = Normalize_{\theta,N}\left(Pool_{\theta,P}\left(Threshold_{\theta,T}\left(Filter_{\theta,F}\left(input\right)\right)\right)\right)$$

where the various θ,X describe parameters for each of the constituent operations. In short, the training process starts with an inclusive family of hypotheses for the cortical computations that could be in one layer – a "space of details" parameterizing neural computations of limited overall complexity. This family of models has yielded success in describing processing across visual cortical areas, and visual encoding more generally. Following basic principles, this underlying feature approach has been demonstrated to achieve good performance for face recognition tasks.[14] For each face in the watch-list considered in this work, four different feature vectors for all four frontalization bins are generated.

Matching

The matching process consists of yaw estimation, bin assignment, frontalization, and feature matching. After a probe image is submitted as input to the system, the face is detected and its yaw angle is calculated using Zhu and Ramanan.[13] Using the estimated yaw angle, the face is assigned to a specific pose bin using Table 1. The face is frontalized using the modified version of Hassner et al.,[2] and the biologically-inspired network is used to extract features. Performance of 1-to-*N* matching then takes place within its respective watch-list matching bin.

Table 1. Pose normalization rules.

Yaw angle	Matching bin
> 45°	Hard Symmetry 1
45° to 15°	Soft Symmetry
15° to -15°	No Symmetry
-15° to -45°	Soft Symmetry
< -45°	Hard Symmetry 2

A "match vector" for each subsequent pair is then calculated using the rule:

$$MatchVector_{ijk} = \left(\left|FeatureVector_{ik} - FeatureVector_{jk}\right|\right)$$

A linear Support Vector Machine (SVM) is then used to calculate a similarity score from any given $MatchVector_{ij}$. The SVM model is trained on the ICB-RW training set, using all true matches and an equal amount of sub-sampled non-matches. The biologically-inspired network is trained on the LFW[15] data set without frontalization, a completely disjoint dataset from ICB-RW using frontalization. Therefore, the SVM learns a decision boundary based on the feature space

of frontalized faces. This way, a more generalizable neural network model is used to generate features, while the decision boundaries are learned for the specific recognition task at hand. Using the trained SVM model, match scores for each feature vector pair are calculated and used to populate a similarity matrix for performance analysis. The entire pipeline is shown in Figure 1.

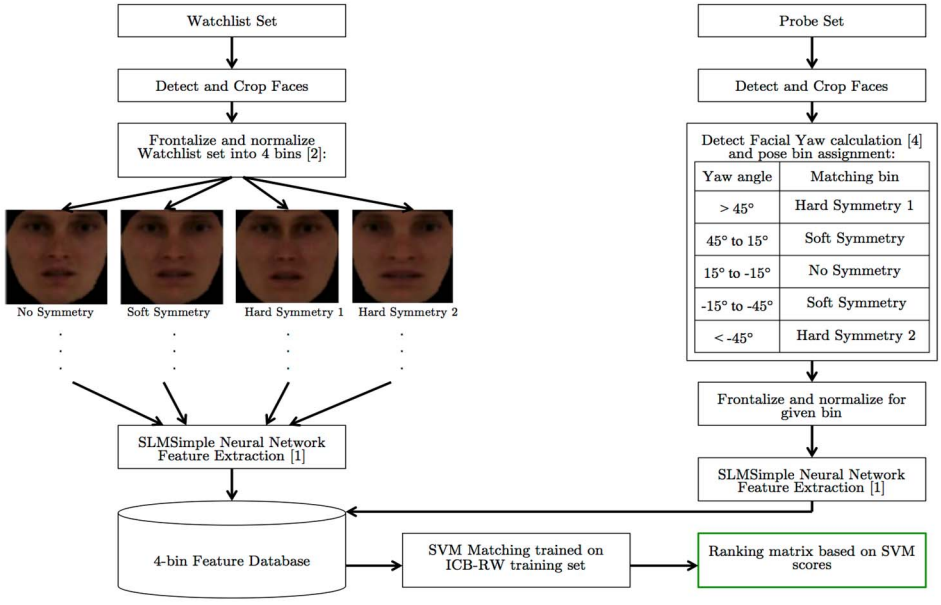

Figure 1. Training and matching pipeline for the proposed algorithm. Four versions of each gallery image are generated using frontalization, and one is placed in each bin. Each image per bin is encoded as a feature vector using the biologically-inspired deep neural network. In the matching phase, the probe's pose is estimated and is assigned to one of the bins, where it is compared with all the feature vectors of that bin, yielding *N* comparison vectors for *N* gallery subjects. These vectors are then fed to a SVM, responsible for verifying if each vector corresponds to a positive or negative pair.

Results and Conclusions

This method produces a high rank-1 match rate. Table 2 shows match rates and the area under the curve of our generated Cumulative Match Characteristic. While rank-1 and 5 show relatively poor performance, the global AUC shows that this DNN+SVM method provides decent discrimination between faces when enough faces are taken into consideration.

Table 2. Rank match rates.

Rank 1	Rank 5	AUC
11.60%	30.40%	0.76%

This method deviates from other proposed methods in a few key ways: it is the only method that uses frontalization to align and normalize the pose of faces. Additionally, the frontalization performs a "smart" binning operation. This allows for a smart feature matching process that can account for extreme pose mismatch. Lastly, this method only incorporates the ICB-RW training data set at the last step within an SVM model, while the biologically-inspired neural network was trained on LFW. This helps insure that the system has not overfit the task at hand, and may help explain why rank-1 and 5 rates are lower than some of the neural network approaches in this competition.

The source code for this work is made publicly available at: https://github.com/joelb92/Smart-Pose-Facial-Matching

EXPLORING FACIAL REGIONS IN UNCONSTRAINED SCENARIOS: EXPERIENCE ON ICB-RW

Ester Gonzalez-Sosa
Ruben Vera-Rodriguez
Julian Fierrez
Javier Ortega-Garcia
Universidad Autonoma de Madrid

Previous works have studied the potential of using facial regions instead of the whole face in biometrics for unconstrained scenarios.[16-17] In Bonnen, Klare, and Jain,[16] four facial regions (eyebrows, eyes, nose, and mouth) were used, conducting some of the experiments with the ARFace, a database with fixed occlusions in a constrained scenario (high resolution with controlled illumination and pose). In Tome et al.,[17] additional face regions were considered (up to 15) using the SCFace database. This database simulates a forensic scenario, including mugshot and CCTV images. This database, though, is not completely realistic, as users cooperate with the system (controlled pose) and the illumination is also controlled. In the present work, we explore face recognition through facial regions on the QUIS-CAMPI dataset. This database is one of the most challenging forensic databases in the literature, as it comprises mugshot images and CCTV images acquired in fully unconstrained scenarios without any cooperation from the users. The CCTV images have variations in pose, occlusions, illumination, distance, expression, etc. Please notice that our intention in this work is not to beat state-of-the-art approaches, but to give an insight into the potential use of facial regions in unconstrained scenarios. Our main objective in this line of work is to devise general face methods to exploit region-based face processing applicable to existing matches. We really believe this region-based processing will benefit even the most advanced face recognition approaches (e.g., based on deep learning) when confronted by challenging scenarios such as the one represented in the ICB-RW competition. With this vision in mind, the present work presents an example of how a robust face matcher based on SIFT can be improved by also considering frontalized facial regions.

Preprocessing, Feature Extraction, and Matching

Preprocessing

Figure 1 shows the general scheme followed in this work. The face is detected using the bounding box information provided as metadata by ICB-RW organizers. The preprocessing stage involves grayscaling, illumination normalization,[18] and resizing (320 x 320). As facial region extraction highly depends on the subject pose, we frontalize the face using the software provided by Hassner et al.[2] The frontalization process involves the estimation of a projection matrix between a query image and a standard 3D reference.

Feature Extraction

Two different features are computed for each image: (i) local binary patterns (LBP) of 9 facial regions and (ii) scale invariant feature transform descriptors (SIFT) of the whole face.

1. Local Binary Patterns of Facial Regions (LBP): In this work, 9 facial regions are extracted from the frontalized face: right eye, left eye, left eyebrow, right eyebrow, nose, mouth, chin, eyes, eyebrows, and face. First, a set of 68 landmarks are extracted through active shape modeling (ASM). Each facial region is extracted from the location of some landmark points as described in Gonzalez-Sosa et al.[19] Then, the facial region is divided into 10 x 10 blocks. The histogram of LBPs (59 uniform patterns) is

computed per each block. The final feature vector of a facial region is the concatenation of the different histograms of LBP computed per block.

2. Scale Invariant Feature Transform Descriptors (SIFT): While local binary patterns highly depend on the spatial correlation between images, SIFT features are more robust against changes in scale and rotations; therefore, they may be more suitable for comparing images without frontalization. In our implementation, SIFT descriptors are computed using cells of 6 x 6 pixels around keypoints and 16 orientations.

Matching

For SIFT descriptors, the similarity between two single images is defined as the number of matched keypoints between the two images, given a certain threshold. The dissimilarity between two LBP descriptors of two facial regions is computed using the Euclidean distance, followed by a normalization by the dimension of the particular facial region feature, to assure that all facial regions contribute similarly.

Experimental Protocol

The QUIS-CAMPI training set is composed of 3 mugshot images and 5 CCTV images per user. In our submitted approach, we only use the frontal mugshot image and the 5 CCTV images as the training images of a particular watch-list subject. At the evaluation phase, we have a test CCTV image, which is preprocessed and frontalized as described earlier. We apply one to one comparisons between the test CCTV image and all the training images belonging to the same watch-list subject before estimating the final score. If frontalization succeeds, these comparisons are carried out using LBP descriptors extracted from 9 facial regions; SIFT descriptors are used otherwise. The final score between a test CCTV image and a watch-list subject derives from the combination of the individual scores that result from the comparisons of the test CCTV image with each of the training images. This combination function depends on the specific face recognition system employed:

1. SIFT-based system: The final similarity score is the maximum of the 6 individual similarities.
2. Frontalized Region-based system: When attempting to compute a final similarity score, we address a N x 9 matrix of similarities, where N is the number of training images from a particular watch-list subject that have been successfully frontalized, and 9 is the number of facial regions considered in each individual comparison. The final score is the sum of the best 5 facial region similarities, having previously chosen the maximum similarity of each facial region.

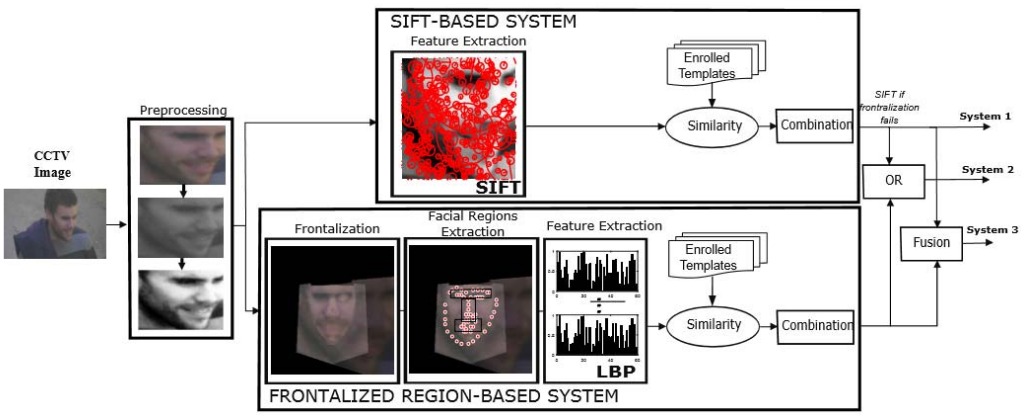

Figure 1. General scheme of the different systems considered in this work: system 1 (baseline), system 2 (submitted) to the ICB-RW Competition, and system 3 (improved).

Results

Results are reported in terms of identification task with rank-1, rank-5, and Area Under the Curve (AUC) between 0 and 1 for the QUIS-CAMPI dataset. Figure 2 shows the cumulative match characteristic curves for the three different systems considered:

- System 1 (baseline): Using only SIFT descriptors (R1=20.0; R5=34.0, AUC=0.69).
- System 2 (submitted): Based on LBP facial regions or SIFT descriptors, depending on the frontalization (R1=24.0; R5=39.1, AUC=0.73).
- System 3 (improved): Based on the fusion of SIFT descriptors and LBP facial regions or only SIFT descriptors, depending on the frontalization (R1=34.2; R5=48.6, AUC=0.80).

The submitted approach improves the baseline system from 0.69 to 0.73 in terms of AUC, and also improves rank-1 and rank-5 rates. The frontalization and the possibility of using similarities of facial regions coming from different training images of the subject may be the reason for this improvement. A big performance improvement is seen with the improvised fusion in which an AUC of 0.80 is obtained, yielding a 15.94 percent relative improvement with respect to the baseline system. Concerning rank-1 rates, there is an absolute improvement of 14.2 with respect to the baseline system. This is due to the complementary information coming from the fusion of SIFT descriptors and the LBP facial regions (when frontalization is possible).

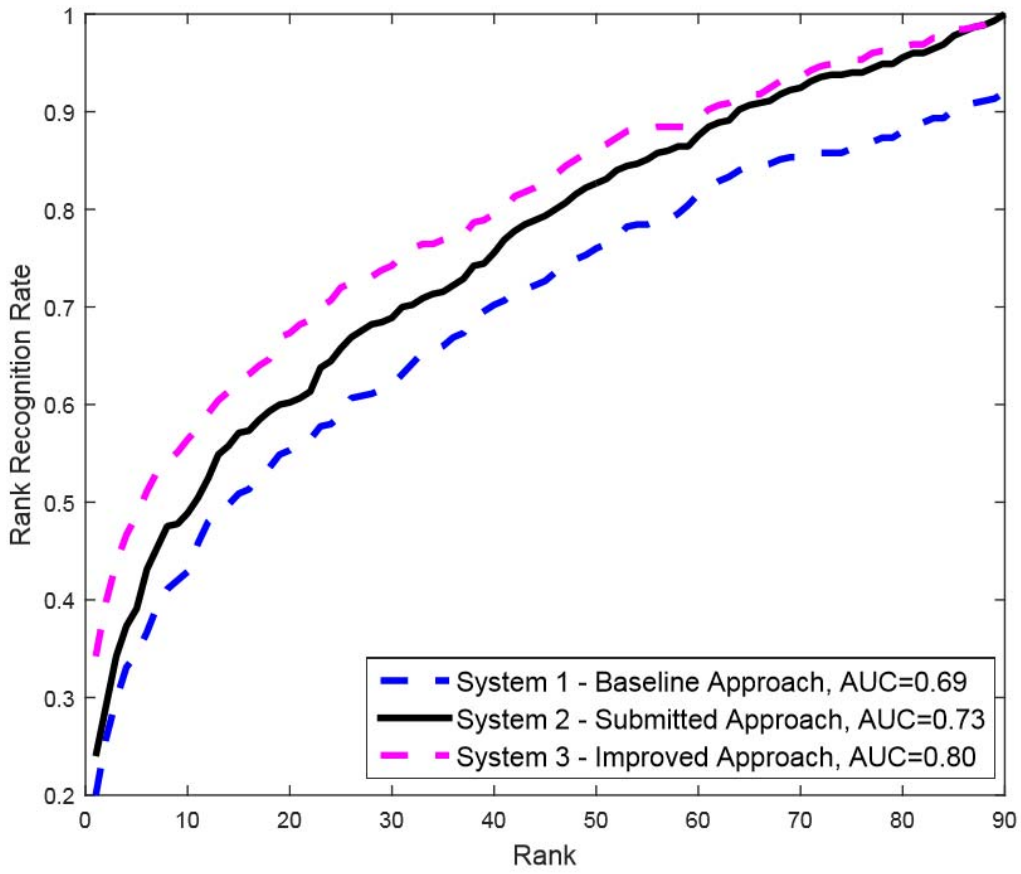

Figure 2. Cumulative Match Characteristic curves for system 1 (baseline), system 2 (submitted), and system 3 (improved).

Conclusion

This work explores the problem of face recognition in real unconstrained scenarios using a facial region approach. Our approach aims to be robust against challenging scenarios, either by using descriptors robust to rotations and changes in scales, or using texture information from different facial regions extracted from a frontalized face. It also introduces a combination function to estimate the best final score among a test CCTV and the training images. Finally, we propose an improved system based on the combination of complementary information coming from SIFT and LBP descriptors that outperformed significantly the submitted approach.

Acknowledgments

This work has been partially supported by project CogniMetrics TEC2015-70627-R (MINECO/FEDER). E. GonzalezSosa is supported by a PhD scholarship from Universidad Autonoma de Madrid.

UNSUPERVISED FACE RECOGNITION IN THE WILD

Michele Nappi
University of Salerno

Daniel Riccio
University of Naples Federico II

Luigi de Maio
Biometric and Imaging Processing Laboratory (BIPLab)

The soaring number of video surveillance cameras that are installed in public places makes face recognition from video surveillance an increasingly important task. In this contribution we present a new unsupervised face identification framework that searches faces extracted from video frames among a set of enrolled identities, which represent a gallery of known persons.

Face recognition from video is attracting ever increasing attention from both academic laboratories and industries, due to its high potential in many real world security applications. Most of the present methods deal with face recognition from videos that supply face images with high-resolution and favorable conditions in terms of pose and illumination.

This scenario is quite far from that, characterized by real video-surveillance applications, where low resolution cameras acquiring unaware people often provide low-quality face images, which are affected by large distortions in terms of non-frontal pose and/or uneven illumination. The main goal of researchers in this field is filling this gap.

As classifying faces acquired in uncontrolled settings is a complex task, most of the present methods are supervised approaches. They rely on a preliminary training stage on labeled faces to learn the structure of the feature space aiming to optimize the separation among different classes.

However, unsupervised methods show the advantage of classifying faces without any previous knowledge of the class distribution. This represents a desirable property when dealing with a large number of clusters with little labeled data.

This contribution proposes a complete framework, namely Unsupervised Face Recognition in the Wild (UFRW) for face recognition in video-surveillance applications, where few pictures per person are provided as enrolled identities that must be identified in single video frames that are submitted to the system as probes.

The UFRW biometric system has been tested in the ICB-RW 2016 challenge, where the goal was to identify persons appearing in video-surveillance frames (still images). Objects to be identified were also provided with high quality images that have been used for enrolling them into the system.

The whole pipeline of UFRW is shown in Figure 1.

Face Detection and Normalization

Many approaches for face detection and normalization have been proposed to achieve invariance to data conditions and to allow biometrics to operate in uncontrolled settings. Though such differences do not hinder the human ability to recognize a person, they can heavily degrade the performance of an automatic recognition system. In order to cope with this problem, UFRW implements both these tasks.

Face Detection

Face detection is the first operation performed by the system, when an image is submitted to UFRW as input. When more than one face appears in the image, the system locates the largest one, as it guesses that the closer the subject to the acquisition camera, the higher the probability of obtaining an accurate identification response. The region of interest (ROI) including the face is located by means of the Viola-Jones[20] global face detector. The extracted ROI then undergoes a further localization process, which searches for 68 facial landmarks. The landmark localization is performed by minimizing a global distance between candidate facial points and their homologues on a general shape model as defined in Grgic, Delac, and Grgic.[21] The shape model is pre-computed over a wide set of annotated training images. The accuracy of the facial landmark localization process represents a critical aspect, as it affects subsequent steps in the recognition pipeline.

Face Normalization

Many algorithms in literature address pose normalization, aiming at improving classification accuracy. The computational cost is the true limit of most of all these approaches when processing a high number of faces. UFRW does not implement any pose correction strategy. However, by exploiting facial landmarks provided by the face detection module, it performs simple preprocessing to adjust the face ROI before inputting it to the face matching module. First of all, the center of the eyes is used to correct head rolling and to compute the inter-ocular distance d (in pixels). The face ROI is then scaled to a factor f=48/d to ensure that the inter-ocular distance of all face images is fixed to 48 pixels. At last, the face ROI is cropped to a fixed size of 64 x 100 pixels. Having obtained the face ROI, it then undergoes an illumination correction process that is performed by applying the weberfaces technique described in Wang et al.[22] This algorithm applies a localized histogram equalization process by exploiting Weber's law, so that the contrast of a pixel is enhanced according to its surrounding luminance.

Face Matching

Video-surveillance applications run in partially or totally uncontrolled settings, providing images that are heavily affected by severe distortions. Since most of the well-known face classification techniques are not robust enough to cope with pose/illumination changes, occlusions and expression variations, they cannot be profitably used in uncontrolled conditions. In this contribution we propose to perform the matching by implementing a localized version of the spatial correlation index. A global correlation index between two images A and B is defined as:

$$s(A,B)=\frac{\sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\left(A(i,j)-\bar{A}\right)\left(B(i,j)-\bar{B}\right)}{\sqrt{\sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\left(A(i,j)-\bar{A}\right)^2\sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\left(B(i,j)-\bar{B}\right)^2}}$$

Where $\bar{A}$ and $\bar{B}$ represent the average values of pixels of A and B, respectively. In UFRW, A and B are partitioned in subregions r_A and r_B, so the correlation values $s(r_A, r_B)$ are computed locally and then cumulated over all subregions. In more detail, for each subregion r_A, UFRW searches in a limited window around the same position in B, the region r_B, which maximizes the correlation $s(r_A, r_B)$. The global correlation $S(A,B)$ is obtained as the sum of the local maxima. Even if this approach is more computationally expensive, it turns out to be more accurate. There

are two considerations motivating this choice. First of all, a higher accuracy is a more stringent requirement in video-surveillance applications when performing off-line video analytics. Secondly, the pre-computation of some quantities in the matching formulae, the code optimization, and the reduced required resolution, allow for the performance of a considerable amount of matches (hundreds) in less than one second on medium-low band computing equipment.

Face Similarity Computation

The identification protocol implemented by UFRW assumes that the system gallery G contains at least one acquisition (template) for each enrolled identity I_j, j=1,...,$|H|$, where H is the set of such identities. A query image p submitted to UFRW is matched against all the templates g_k, k=1,...,$|G|$, in the gallery G by computing the corresponding correlation index $S(p, g_k)$. The resulting list of scores is sorted in decreasing order, and the identity I_j with the highest correlation is returned.

Results

The UFRW biometric system was evaluated on face images acquired in uncontrolled conditions. In particular, two different datasets have been considered. The former is SCFace,[21] which includes 130 subjects who have been captured at three different distances, with eight devices (cam1, cam2,…,cam8), five of which were in visible daylight, two during night in visible light, and one in infrared. In this experiment we only considered visible daylight images which were captured by the first device (cam1).

We also tested UFRW on images provided in the QUIS-CAMPI database.[23] In this experiment, 90 out of the 320 subjects have been considered for testing. Faces extracted from high-quality registration images (frontal view) have been used to enroll subjects into the gallery G, while 450 images of the same subjects have been included in the probe set P. Probe images have been automatically acquired by a parent-child surveillance system during the QUIS-CAMPI project.

Table 1. Performance of UFRW in terms of Cumulative Match Score (rank-1 and rank-2) and Area Under CMC Curve (AUC).

Dataset	Rank 1 (%)	Rank 5 (%)	AUC (CMC curve)
SCFace	70.4	90.3	0.954
QUIS-CAMPI	11.1	25.1	0.694

From results in Table 1, it can be observed that the recognition performance of UFRW on SCFace is significantly better than that obtained on QUIS-CAMPI. This can be explained by considering that face images in SCFace show less pose distortions than those in QUIS-CAMPI, even if both sets of video frames are extracted from video-surveillance streams. As a matter of fact, UFRW does not implement any pose frontalization stage, so that it encounters difficulties when dealing with large pose distortions. Moreover, UFRW is an unsupervised face recognition system, so it has no previous knowledge of the class distribution of the enrolled subjects. We think this is a desirable property for a face biometric system, as it does not need a retraining stage if new subjects are enrolled into the gallery. In other words, new identities can be added to or deleted from the gallery without stopping and restarting the system.

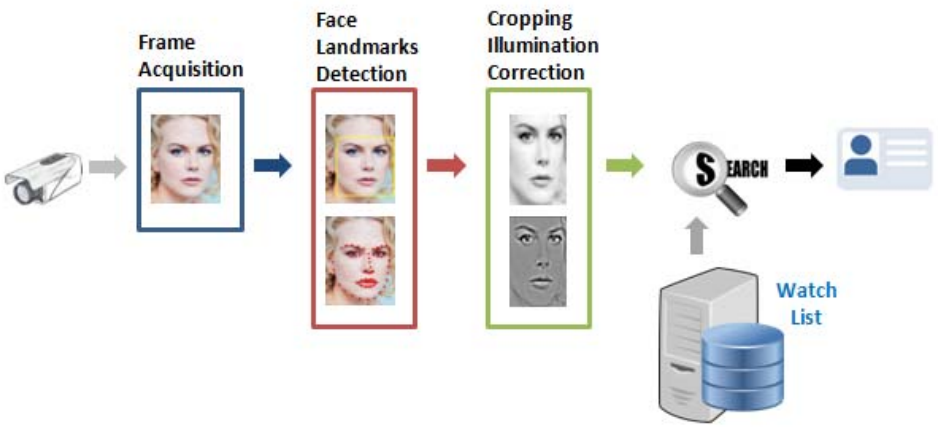

Figure 1. The system architecture of UFRW.

Conclusions

In this contribution, we presented a new face recognition system, namely UFRW, for video-surveillance applications. It is an unsupervised framework, which automatically detects and normalizes the face region from still video frames and searches for a match in a gallery of enrolled subjects according to a localized reformulation of the correlation index. Preliminary results show that there is room for improvement in terms of recognition performance. In future work, we will include a pose normalization process in the system pipeline in order to mitigate the effect of pose distortions.

Acknowledgments

This work was partially supported by the BSCube s.r.1.

REFERENCES

1. H. Gao, H.K. Ekenel, and R. Stiefelhagen, "Combining View-based Pose Normalization and Feature Transform for Cross-Pose Face Recognition," *International Conference on Biometrics* (ICB), 2015.
2. T. Hassner et al., "Effective Face Frontalization in Unconstrained Images," *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2015, pp. 4295–4304.
3. V. Kazemi and J. Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees," *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2014.
4. O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," *British Macine Vision Conference* (BMVC), 2015.
5. K. Simonyan et al., "Fisher Vector Faces in the Wild," *British Machine Vision Conference* (BMVC), 2013.
6. J.C. Neves et al., "QUIS-CAMPI: Extending in the wild biometric recognition to surveillance environments," *International Conference on Image Analysis and Processing* (ICAP), 2015, pp. 59–68.
7. D. Bolme et al., "The CSU face identification evaluation system: Its purpose, features, and structure," *Computer Vision Systems*, Springer, 2003.
8. J. Phillips and A. O'Toole, "Comparison of human and computer performance across face recognition experiments," *Image and Vision Computing*, vol. 32, no. 1, 2014, pp. 74–85.

9. D. Yi et al., "Learning face representation from scratch," arXiv preprint, 2014; arXiv:1411.7923.
10. R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," *IEEE International Conference on Computer Vision and Pattern Recognition* (CVPR), 2006.
11. X. Zhu et al., "Face Alignment Across Large Poses: A 3D Solution," arXiv preprint, 2015; arXiv:1511.07212.
12. J. Yan et al., "Learn to combine multiple hypotheses for accurate face alignment," *IEEE International Conference on Computer Vision Workshops*, 2013.
13. X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2012.
14. D. Cox and N. Pinto, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops* (FG 2011), 2011.
15. G.B. Huang et al., *Labeled faces in the wild: A database for studying face recognition in unconstrained environment*, technical report 1(2):07-49, University of Massachusetts, Amherst, 2007.
16. K. Bonnen, B.F. Klare, and A.K. Jain, "Component-based representation in automated face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, 2013, pp. 239–253.
17. P. Tome et al., "Combination of face regions in forensic scenarios," *Journal of Forensic Sciences*, vol. 60, no. 4, 2015, pp. 1046–1051.
18. V. Štruc and N. Pavešic, "Photometric normalization techniques for illumination invariance," *Advances in Face Image Analysis: Techniques and Technologies*, Y.-J. Zhang, IGI Global, 2011.
19. E. Gonzalez-Sosa et al., "Dealing with occlusions in face recognition by region-based fusion," *Proceedings of the International Conference on Security Technology* (ICCST), 2016.
20. P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2001, pp. 511–518.
21. M. Grgic, K. Delac, and S. Grgic, "SCface – Surveillance cameras face database," *Multimedia Tools Applications Journals*, vol. 51, no. 3, 2009, pp. 863–879.
22. B. Wang et al., "Illumination normalization based on Weber's law with application to face recognition," *IEEE Signal Processing Letters*, vol. 18, no. 8, 2011, pp. 462–465.
23. J. Neves and H. Proenca, *QUIS-CAMPI Multi-Biometrics dataset*, 5 May 2016; http://quiscampi.di.ubi.pt/.