# Deep Face Recognition for Surveillance Applications

Klemen Grm, *Student Member, IEEE,* and Vitomir Struc, *Member, IEEE*

***Abstract*—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.**

***Index Terms*—Surveillance, face recognition, deep models, Quis-campi**

## I. Introduction

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 pretrained VGG network to process the security footage and extract high-level facial representations. A softmax classifier is then trained on top of the very 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 inference in the target domain (i.e., on the security footage). The presented approach was submitted to the *International Challenge on Biometric Recognition in the Wild* (ICB-RW) organized in the scope of the International Conference on Biometrics 2016 and ranked second among nine participants.

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 [3] 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.

Klemen Grm and Vitomir Struc are with the University of Ljubljana, Slovenia, e-mail: klemen.grm,vitomir.struc@fe.uni-lj.si.

## II. Deep learning for surveillance applications

### A. 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 performance 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 $\theta = \{\mathbf{W}\}$ that are learnt 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:

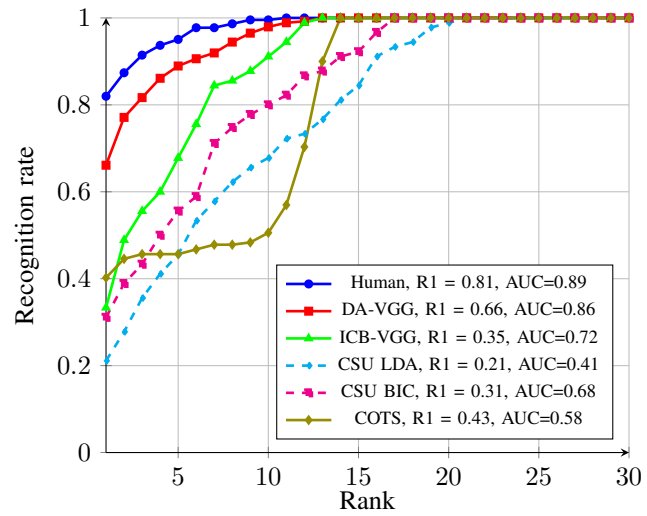
$$\mathbf{y}_n = f_{\theta_n}(\mathbf{y}_{n-1}) = \sigma(\mathbf{y}_{n-1} * \mathbf{W}_n), \tag{1}$$

where $\mathbf{y}_n$ and $\mathbf{y}_{n-1}$ $(1 \leq n \leq N)$ represent the outputs of $n$-th and $(n-1)$-th layer, respectively, $\sigma$ 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., $\theta_n = \{\mathbf{W}_n\}$, and the input to the first layer ($n = 1$) are the raw (unprocessed) images. An $N$-layer deep CNN is then described as:

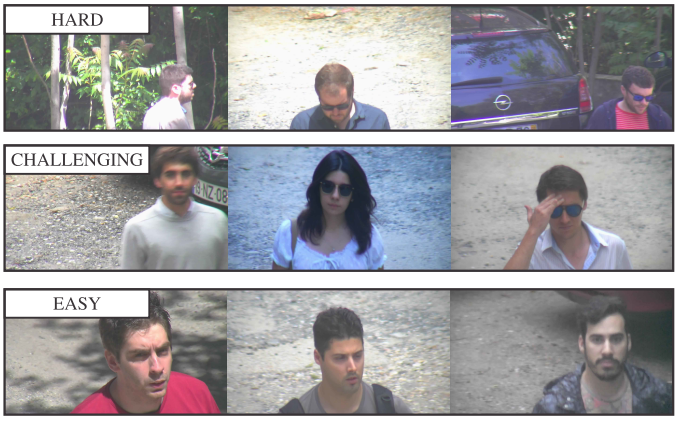
$$\mathbf{y} = (f_{\theta_N} \circ f_{\theta_{N-1}} \circ ... \circ f_{\theta_1})(\mathbf{x}), \tag{2}$$

where $\mathbf{x}$ and $\mathbf{y}$ 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.

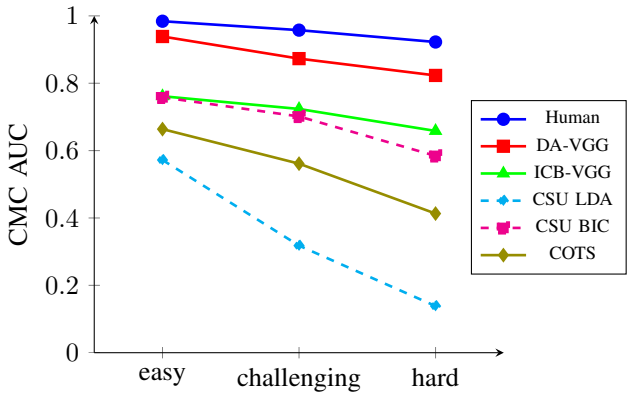
By training convolutional networks via gradient descent, the image representation is learned directly from the input data in an end-to-end manner, as opposed to classical computer vision approaches where the image descriptors are typically hand-crafted before being fed to some classifier.



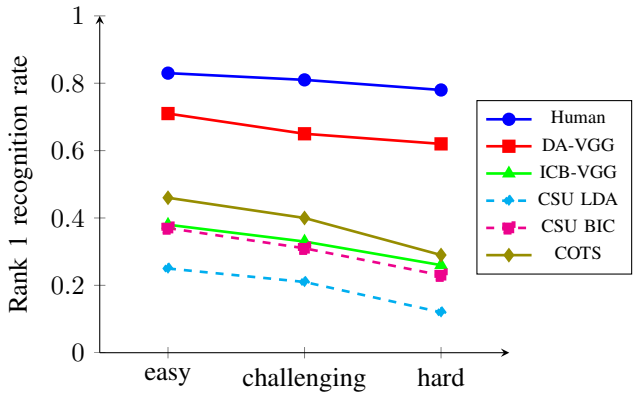

(a) CMC curves for all tested methods




(b) Sample images from the ICB-RW dataset




(c) AUC values across three levels of difficulty




(d) R1 values across three levels of difficulty

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 figure is best viewed in color.

### *B. The VGG architecture*

The VGG network architecture, introduced for face recognition in [2], 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 architectures, i.e.,:

- **Small filters**: All convolutional filters are of size $3 \times 3$ pixels, as opposed to earlier CNNs which used much larger filter sizes. By using multiple $3 \times 3$ 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 $2 \times 2$ 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.

### *C. 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 $2.6 \times 10^6$ facial images over several weeks on a computer equipped with 4 high-performance GPUs [2]. To make the VGG network applicable to surveillance scenarios, we resort to domain adaptation techniques and apply them to the pretrained VGG (face) convolutional network from [2]. We perform net surgery on the VGG network and use the pretrained configuration

for representation calculation. On top of the network we train a probabilistic multi-class softmax classifier using the development set of the ICB-RW data.

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

$$\mathbf{p} = \frac{e^{\mathbf{W}^T\mathbf{y}+\mathbf{b}}}{\sum_{i=0}^{M} e^{\mathbf{w}_i^T\mathbf{y}+b_i}} \tag{3}$$

where the image representation $\mathbf{y} \in \mathbb{R}^{K\times 1}$ is generated by the pretrained VGG network and the matrix $\mathbf{W} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \ldots, \mathbf{w}_M^T]^T \in \mathbb{R}^{K\times M}$ and the vector $\mathbf{b} = [b_1, b_2, \ldots, b_M]^T \in \mathbb{R}^{M\times 1}$ are learned parameters of the classifier. The classifier is trained via minibatch error back-propagation 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 $\mathbf{y}$ is classified into the class with the highest posterior probability.

With the presented approach the pretrained 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.

## III. Experiments and results

We assess the suitability of the domain-adapted VGG network (DA-VGG) for surveillance scenarios on part of the Quis-Campi [3] dataset used for the ICB-RW 2016 competition. The data contains gallery and probe images of 90 distinct subjects (see Fig. 1 (b)). 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, while 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 (CMCw), the rank-1 (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 $224 \times 224$ pixels.

We provide comparative 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) [4], a deep convolutional neural network based on the VGG architecture trained solely on the development set of the 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 [5].

The CMC plots of the experiments are presented in Fig. 1 (a). The DA-VGG network outperforms the CSU baselines with a margin of over 30% in terms of the rank-1 recognition rate. The DA-VGG network also results in better performance than the ICB-VGG network, 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%, which is below the 66% ensured 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% on this dataset at rank 1 in favor of humans. This observation is in line with previous work, e.g. [5], 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 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 Fig. 1(b). The AUC values and rank 1 recognition rates across the three levels are shown in Figs. 1(c) and (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 "easy" images, while the performance gap is bigger for the "hard" images.

## IV. 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.

## References

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

[2] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.

[3] J.C. Neves, G. Santos, S. Filipe, E. Grancho, S. Barra, F. Narducci, and H. Proença, "Quis-campi: Extending in the wild biometric recognition to surveillance environments," in *ICIAP*. Springer, 2015, pp. 59–68.
[4] D. Bolme, R. Beveridge, M. Teixeira, and B. Draper, "The CSU face identification evaluation system: its purpose, features, and structure," in *Computer Vision Systems*, pp. 304–313. Springer, 2003.
[5] J. Phillips and A. O'Toole, "Comparison of human and computer performance across face recognition experiments," *IMAVIS*, vol. 32, no. 1, pp. 74–85, 2014.