Designing a Low-Resolution Face Recognition System for Long-Range Surveillance

Yuxi Peng, Luuk Spreeuwers and Raymond Veldhuis
Faculty of Electrical Engineering, Mathematics and Computer Science, SCS researchgroup,
Univerisity of Twente, Enschede, the Netherlands
Email:{y.peng, l.j.spreeuwers, r.n.j.veldhuis}@utwente.nl

***Abstract*—Most face recognition systems deal well with high-resolution facial images, but perform much worse on low-resolution facial images. In low-resolution face recognition, there is a specific but realistic surveillance scenario: a surveillance camera monitoring a large area. In this scenario, usually the gallery images are of high-resolution and the probe images are of various low-resolutions depending on the distances between the subject and the camera. In this paper, we design a low-resolution face recognition system for this scenario. We use a state-of-the-art mixed-resolution classifier to deal with the resolution mismatch between the gallery and probe images. We also set up experiments to explore the best training configuration for probe images of various resolutions. Our experimental results show that one classifier which is trained on images of various resolutions covering the whole range has promising results in the long-range surveillance scenario. This system has at least as good performance as combining multiple face recognition systems that are optimised for different resolutions.**

I. Introduction

Face recognition at a distance is a challenging subject. The face images captured at a distance are of low-resolution so that they contain less information. In addition, the most common application for face recognition at a distance is camera surveillance, where the images are usually captured in uncontrolled situations, which results in illumination and pose variations. The images are also often noisy due to low light or compression artifacts. Those various problems suffered by the images in low-resolution face recognition make them more difficult to recognize than those in high-resolution face recognition. In this paper, we will focus on the low-resolution problem.

There are many methods that have been developed for low-resolution face recognition. Some of the methods improve face recognition performance on low-resolution images by applying super-resolution to increase the image resolution. Zhang et al. [1] proposed a super-resolution method in morphable model space, which provides high-resolution information required by both reconstruction and recognition. Zou and Yuen [2] developed a data constraint for reconstructing super-resolution image features so that both the distances between the reconstructed images and the corresponding high-resolution images and the distances between super-resolution images from the same class are minimized. There are also face recognition methods that perform face recognition directly on low-resolution images. Li et al. [3] proposed a method that projects both high-resolution gallery and low-resolution probe to a unified feature space for classification using coupled mappings which minimize the difference between corresponding images. Moutafis and Kakadiaris [4] proposed a method that learns semi-coupled mappings for optimized representations. The mappings aim at increasing class-separation for high-resolution images and mapping low-resolution images to their corresponding class-separated high-resolution data. Peng et al. [5] proposed a likelihood ratio based method for direct comparison between images of different resolutions. All these works consider a specific resolution.

In face recognition at a distance, if the camera is monitoring a large area, the face images captured at a long distance can have very different resolutions. There are two approaches to deal with this situation. The first one is to improve acquisition devices so that the images captured at the farthest distance have a high enough resolution for recognition. An acquisition system using wide-field-of-view cameras and near-field-of-view cameras is proposed by Wheeler et al. in [6]. Wide-field-of-view cameras monitor the large area which detect and locate the person. Then Near-field-of-view cameras are controlled automatically to capture high-resolution face images. Commercial face recognition system is used for recognition. This system can detect faces at distances of 25-50 m and recognize faces (first successful face recognition) at distances of 15-20 m. This type of camera systems is also used in [7], [8]. The second approach is to design a face recognition system for images captured at various distances. However, there is only a limited number of publications on this topic. Moon and Pan [9] proposed an LDA-based long distance face recognition algorithm. They demonstrated that using images from multiple distances for training has better recognition performance on far distances than using single near distance images for training. Tome et al. [10] proposed an estimator of the acquisition distance based on the segmented face area and the full image area. In the face verification process, the scores of DCT-GMM-based system and PCA-SVM-based system are fused based on the distance estimator so that for far distance the DCT-GMM-based system has more weight.

We follow the second approach and explore what is important in designing a low-resolution face recognition system for long range surveillance. We use a state-of-the-art low-resolution face recognition method and explore how we should set up the classifiers to cover the whole range in this scenario.

The remainder of this paper is organised as follows: we

describe the scenario of the problem and propose the hypothesis in Section II. The face recognition method we use in the experiments is introduced in Section III. Experimental results are reported in Section IV. Section V gives conclusions.

II. Scenario and hypothesis

We focus on this specific but common surveillance circumstance: a surveillance camera monitoring a large area. The camera is connected to a face recognition system. When a person appears in the camera view, the face recognition system detects the person's face and compares it with the suspects' faces in the database. The output of the system is the decision whether this person is one of the suspects or not. In this situation, the gallery images (from the suspects) that are of high resolution. The probe images, captured at a distance, are of low resolution. Because the person can show up at any point of the monitored area, the resolution of the probe images can vary a lot.

To design a face recognition system for this scenario, we address two problems. Firstly, there is a resolution mismatch between high-resolution gallery and low-resolution probe while most classifiers are designed for high-resolution images and can only work properly for images of the same resolution. Secondly, usually a single classifier is not able to achieve sufficient recognition performance for images captured at very different distances.

To solve the first problem, we use the MixRes classifier [5]. After training with high-resolution and low-resolution image pairs, the MixRes classifier can directly compare low-resolution probe to high-resolution gallery. It is shown in [5] that this method has promising performance on very low-resolution probes.

To address the second problem, we test the following hypothesis: it is beneficial for the recognition performance of long range face recognition to combine several classifiers that are tuned to images of different resolutions. Each of the classifiers gives the best face recognition performance on images of a certain resolution. The combination of those classifiers are supposed to give optimal results across the whole range of distances.

III. MixRes classifier

Here we briefly describe the MixRes classifier, which is originally named Mixed-Resolution Biometric Comparison [5]. The MixRes classifier is especially designed for comparing images captured at different distances. It is based on the likelihood ratio of a pair of mixed-resolution input images. This method is similar to [11] (also described in [12]) which is derived for comparing images of the same resolution.

This classifier transforms the reference sample x and test sample y to a common lower dimensional space by

$$\mathbf{x}_\mathrm{c} = \mathbf{Z}_\mathrm{R}\left(\mathbf{x} - \bar{\mathbf{r}}\right), \tag{1}$$

$$\mathbf{y}_\mathrm{c} = \mathbf{Z}_\mathrm{T}\left(\mathbf{y} - \bar{\mathbf{t}}\right), \tag{2}$$

where $\bar{\mathbf{r}}$ and $\bar{\mathbf{t}}$ are the grand means of, respectively, the reference and probe training sets. The transformations $\mathbf{Z}_\mathrm{R}$ and $\mathbf{Z}_\mathrm{T}$ are computed in a training phase. They are built up of a PCA step, reducing the dimensionality of probe and reference to manageable proportions as well as whitening them, followed by an LDA step that aims for optimal discrimination after transformation to a common lower dimensional subspace. This method and the training procedure are described in [5]. The log-likelihood ratio is then computed as

$$\log(\mathrm{l}(\mathbf{x}_\mathrm{c}, \mathbf{y}_\mathrm{c})) = -\frac{1}{2}\sum_{i=1}^{D}\log\left(1-\nu_i^2\right) + \frac{1}{4}\mathrm{s}(\mathbf{x}_\mathrm{c}, \mathbf{y}_\mathrm{c}), \tag{3}$$

where ν_i is the between-class covariance of feature element i after the LDA step and

$$\mathrm{s}(\mathbf{x}_\mathrm{c}, \mathbf{y}_\mathrm{c}) = \tag{4}$$
$$-\sum_{i=1}^{D}\frac{\nu_i}{1-\nu_i}(x_{\mathrm{c},i} - y_{\mathrm{c},i})^2 + \sum_{i=1}^{D}\frac{\nu_i}{1+\nu_i}(x_{\mathrm{c},i} + y_{\mathrm{c},i})^2.$$

IV. Experiments

The goal of our experiment is to test the hypothesis formulated in Section II that combining multiple face recognition systems, optimised for different resolutions improves the face recognition performance in the long-range surveillance scenario.

The Human ID database [13] is chosen because it is suitable for simulating the scenario as described in Section II. This database contains high-resolution mug shots which we use as gallery. It also contains parallel gait videos which are the best source for the probe images. There are maximum four sessions recorded at different time. We have 588 mug shots from 312 subjects. There are at most four mug shots per person. Most people have one or two mug shots. The parallel gait videos are captured when a person was walking towards the camera. Because the videos are not taken under controlled condition, there are some pose and illumination variations. We use Viola-Jones face detector [14] (implemented in MATLAB) to detect the faces in the videos. From the detected faces we choose the images of near frontal pose and relatively good quality for our experiments. From the detected faces we selected images with nine different resolutions: 70×70, 60×60, 50×50, 45×45, 40×40, 35×35, 30×30, 25×25, and 23×23 pixels. The distance between the eyes goes down from 28 pixels for 70×70 to 9 pixels for 23×23. For each resolution, two images are randomly selected from each video. The number of images for each resolution is different because some of the videos do not have images with all the nine resolutions. Detailed information about the data we use is in Table I. All the images are aligned using manually marked eye-coordinates. The regions of interest are cropped using an elliptic mask. Sample images are shown in Fig. 1.

In order to test our hypothesis, we performed experiments, which are discussed below.

First, we test on how the classifier performs on images of different resolutions when it is trained with images of different resolutions. This is not only to test our hypothesis is correct, but also to find out if the difference is significant between the classifiers trained with different resolution images.

TABLE I
NUMBER OF IMAGES AND SUBJECTS OF EACH RESOLUTION USED IN OUR EXPERIMENTS. RES: RESOLUTION. NI: TOTAL NUMBER OF IMAGES. NS: NUMBER OF SUBJECTS.

Res	70	60	50	45	40	35	30	25	23
Ni	707	664	755	837	768	811	827	877	873
Ns	251	259	279	282	281	276	276	276	272

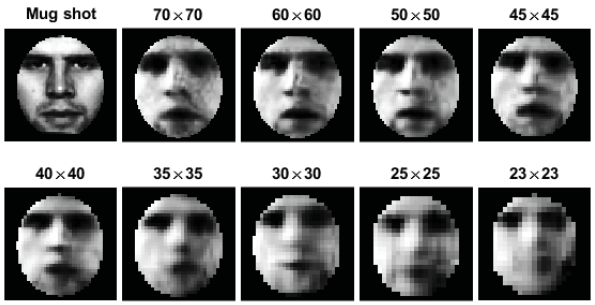

Fig. 1. Sample images of each resolution after pre-processing.

Cross-validation is used in our experiments. During the training procedure, we randomly selected images of 200 subjects for training. The high-resolution training images are from the mug shots and the low-resolution training images are from the video images of each resolution. Because the high-resolution training images are always from the mug shots set, we always mean the low-resolution training sets when we refer to different training sets in the remainder of the paper. The images of the rest of the subjects are used for testing. There is no overlap between subjects for training and testing. After training with images of each resolution, we test the classifier using images of all the nine resolutions. The cross-validation has 100 rounds and the average verification rates (also known as genuine acceptance rates) at False Acceptance Rate (FAR) equals to 0.1 are shown in Fig. 2. The standard deviations of the verification rates are around 0.04.

As we can see, for each probe resolution, training with images of the same resolution gives the best results. Especially when comparing the classifiers trained with images of 70×70 and 23×23 pixel-resolutions, the performance differences are significant: the difference of the verification rates on 70×70 probe is 0.25 and the difference on 23×23 probe is 0.13. The classifiers, which are trained with resolutions different from the probe resolution, usually perform worse at the probe resolution. This supports our hypothesis. We can design a system with nine classifiers, each of them dedicated to a certain probe resolution. However, some of the classifiers, which were trained with images of neighbouring resolutions, have similar performance. For example, the classifiers 70 and 60 has similar performance on probe resolutions 70×70 and 60×60. This means it could be possible to reduce the number of classifiers needed in the system.

In the second experiment we explore how many classifiers

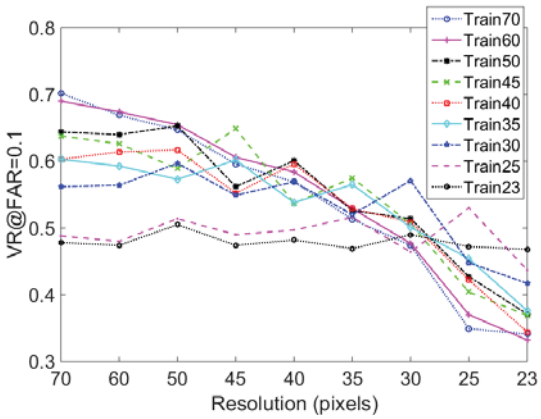

Fig. 2. Verification results of training and testing with images of different resolutions. X axis: probe image resolution, Y axis: Verification Rate (VR) at FAR 0.1.

are necessary in the system. There are four different configurations of classifiers trained with image sets of different resolution divisions, shown in Table II.

TABLE II
DIVISION OF RESOLUTIONS FOR TRAINING IN THE SECOND EXPERIMENT.

Division	Resolutions
DIV1	70, 60, 50, 45, 40, 35, 30, 25, 23
DIV2	70-50, 45-40, 35-30, 25-23
DIV3	60-40, 35-23
DIV4	70-23

DIV1 has nine classifiers trained on each resolution. DIV2 has combined two or three neighbouring resolutions in the training set which results in four classifiers. DIV3 has two classifiers, each of them are trained on images of four neighbouring resolutions. DIV4 only has one classifier, but the training images have all the nine resolutions. All images in each training set are up-sampled to the highest resolution in this training set. For example, the training set of the second classifier in DIV2 consists of images of original resolutions 45×45 and 40×40. The images of original resolution 40×40 are up-sampled to 45×45 before they are used to train the classifier. In the testing phase, if the input image has a resolution of 45×45 or 40×40, they will be scaled to 45×45 and this (the second) classifier is used to compute scores.

To ensure a fair comparison between the four settings, we randomly select five images per subject for training for all the classifiers in the four divisions even though much more images are available in the last three settings. The results are shown in Fig. 3.

As we can see, the performance on each testing resolution is similar for all the four training divisions. The differences between the average values are within their standard deviations. This means that, for a long range surveillance system, it is not necessary to have a number of classifiers optimised for different resolutions. We can use only one classifier but

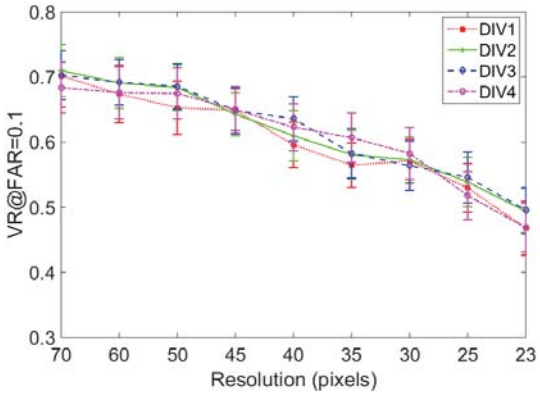

Fig. 3. Verification results of classifiers trained with different resolution divisions of the training data. X axis: probe image resolution, Y axis: Verification Rate (VR) at FAR 0.1.

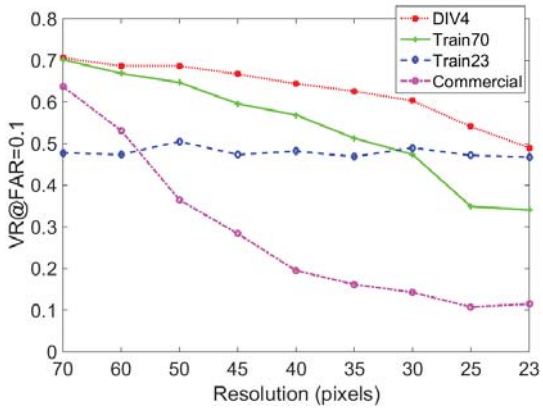

Fig. 4. Comparison of DIV4, Train70 and Train23 to FaceVACS. X axis: probe image resolution, Y axis: Verification Rate (VR) at FAR 0.1.

trained on images captured at different distances. Although the first experiment confirmed our hypothesis and showed that it is indeed beneficial to combine classifiers tuned to various resolutions, the second experiment showed that the same performance can be achieved when one classifier is trained with images with varying resolutions. The latter solution has a lower computational complexity. In addition, because there is usually not that many real low-resolution training images available, a single classifier can make the best use of the training data.

Then we compare the results of our optimised system, which is the DIV4 from the second experiment, to the results of Train70 and Train23 from the first experiment and a commercial face recognition system. We compare with this commercial system because commercial face recognition systems are used in real-life surveillance cases. The results of the commercial system are obtained using all the images available at each resolution (no training image required), and manually marked eye-coordinates of the images are provided to the system. To make best use of the training images in DIV4, 40 images per subjects are used to train the classifier. The other configurations are the same as in the previous experiments. The verification rates at FAR equals to 0.1 are shown in Fig. 4.

As we can see, DIV4 performs the best across all the testing resolutions. Train70 has similar performance as the performance of DIV4 at resolution 70×70, but the difference between them becomes significant when the resolution decreases. This is similar with Train23, which has the closest result to DIV4 at resolution 23×23. The commercial face recognition system performs differently as it is designed for high-resolution face recognition. At the highest resolution 70×70, it performs only slightly worse than DIV4, but the performance dropped quickly when the resolution decreases and from resolution 40×40 to lower, the results are very close to random guessing.

V. Conclusion

Most face recognition systems deal well with high-resolution facial images, but perform much worse on low-resolution facial images. We focus on a specific but realistic surveillance scenario: face recognition at a distance for long range surveillance. Our aim is to design a face recognition system for a range of resolutions. We use an existing mixed-resolution face recognition method and investigated whether it is beneficial for the recognition performance of long range face recognition to combine several classifiers that are tuned to images of different resolutions. Our experimental results show that if a classifier is only trained on images captured at a single distance, it could not perform well on the images from a very different distance. However, if we combine the images captured at various distances for training, a single classifier can perform at least as good as a combination of different classifiers when each of them are trained on images captured at a single distance. We also show that this one classifier system outperforms a state-of-the-art commercial face recognition system.

References

[1] D. Zhang, J. He, and M. Du, "Morphable model space based face super-resolution reconstruction and recognition," *Image and Vision Computing*, vol. 30, pp. 100–108, February 2012.

[2] W. Zou and P. Yuen, "Very low resolution face recognition problem," *Image Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 327–340, Jan 2012.

[3] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *Signal Processing Letters, IEEE*, vol. 17, no. 1, pp. 20 –23, jan. 2010.

[4] P. Moutafis and I. A. Kakadiaris, "Semi-coupled basis and distance metric learning for cross-domain matching: Application to low-resolution face recognition," in *Proc. International Joint Conference on Biometrics*, Clearwater, FL, September 29 - October 2 2014.

[5] Y. Peng, L. J. Spreeuwers, and R. N. J. Veldhuis, "Likelihood ratio based mixed resolution facial comparison," in *3rd International Workshop on Biometrics and Forensics (IWBF2015)*, Gjøvik, Norway, March 2015.

[6] F. W. Wheeler, R. Weiss, and P. H. Tu, "Face recognition at a distance system for surveillance applications," in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, Sept 2010, pp. 1–8.

[7] U. Park, H.-C. Choi, A. Jain, and S.-W. Lee, "Face tracking and recognition at a distance: A coaxial and concentric ptz camera system," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 10, pp. 1665–1677, Oct 2013.

[8] D. Tran, T. Nguyen, H. Bui, S. Nguyen, H. Hoang, T. Pham, and T. de Souza-Daw, "Dual ptz cameras approach for security face detection," in *Communications and Electronics (ICCE), 2014 IEEE Fifth International Conference on*, July 2014, pp. 478–483.

[9] H.-M. Moon and S. B. Pan, "Long distance face recognition for enhanced performance of internet of things service interface," *Computer Science and Information Systems*, vol. 11, pp. 961–974, 2014.

[10] P. Tome, J. Fierrez, F. Alonso-Fernandez, and J. Ortega-Garcia, "Scenario-based score fusion for face recognition at a distance," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, June 2010, pp. 67–73.

[11] R. N. J. Veldhuis and A. M. Bazen, "One-to-template and one-to-one verification in the single- and multi-user case," in *26th Symposium on Information Theory in the Benelux, Brussels, Belgium*, Brussels, May 2005, pp. 39–46.

[12] L. Spreeuwers, "Breaking the 99% barrier: optimisation of three-dimensional face recognition," *IET Biometrics*, vol. 4, no. 3, pp. 169–178, 2015.

[13] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi, "A video database of moving faces and people," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 812–816, May 2005.

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I–511–I–518 vol.1.