Dynamic Amelioration of Resolution Mismatches for Local Feature Based Identity Inference

Yongkang Wong, Conrad Sanderson, Sandra Mau, Brian C. Lovell
NICTA, PO Box 6020, St Lucia, QLD 4067, Australia *
The University of Queensland, School of ITEE, QLD 4072, Australia

Abstract

While existing face recognition systems based on local features are robust to issues such as misalignment, they can exhibit accuracy degradation when comparing images of differing resolutions. This is common in surveillance environments where a gallery of high resolution mugshots is compared to low resolution CCTV probe images, or where the size of a given image is not a reliable indicator of the underlying resolution (e.g. poor optics). To alleviate this degradation, we propose a compensation framework which dynamically chooses the most appropriate face recognition system for a given pair of image resolutions. This framework applies a novel resolution detection method which does not rely on the size of the input images, but instead exploits the sensitivity of local features to resolution using a probabilistic multi-region histogram approach. Experiments on a resolution-modified version of the "Labeled Faces in the Wild" dataset show that the proposed resolution detector frontend obtains a 99% average accuracy in selecting the most appropriate face recognition system, resulting in higher overall face discrimination accuracy (across several resolutions) compared to the individual baseline face recognition systems.

1 Introduction

Face images obtained in surveillance scenarios typically have issues such as misalignment and variations in pose and illumination. Here we address a further issue, namely varying image resolution [9], encountered while undergoing real-world system trials for the UK police and other agencies. Mismatched resolutions between probe and gallery images can cause significant performance degradation for face recognition systems, particularly those which use high-resolution faces (*e.g.* mugshots or passport photos) as gallery images. Another source of resolution mismatches is due to the fact that the size (in terms of pixels) of a given face image may *not* be a reliable indicator of the underlying optical resolution. Examples include: **(i)** poor quality optics in low-cost cameras can act as low-pass filters; **(ii)** poor focus and over-exposure can result in blur and loss of detail; **(iii)** a given gallery face is provided in an already resized form and the original size is unknown (*e.g.* digital scan of a photograph).

Face recognition approaches can be placed into two general families: holistic and local-feature based. In typical holistic approaches, a single feature vector describes the entire face and the spatial relations between face characteristics (*e.g.* eyes) are rigidly kept. Examples of such systems include PCA and Fisherfaces [2]. In contrast, local-feature based approaches describe each face as a set of feature vectors (with each vector describing a small part of the face), with relaxed constraints on the spatial relations between face parts [4]. Examples include systems based on elastic graph matching, hidden Markov models (HMMs) and Gaussian mixture models (GMMs) [4].

Local-feature based approaches have the advantage of being considerably more robust against misalignment (caused by automatic face detectors) as well as variations in illumination and pose [4, 11]. As such, these systems are more suitable for dealing with faces obtained in surveillance contexts. However, almost all of the literature on addressing resolution mismatches (*e.g.* [5, 7]) deals with holistic approaches and naively assumes that faces are localised perfectly (*i.e.* no misalignment) as well as being frontal (*i.e.* no pose variations).

In typical local-feature based face recognition systems, the size of probe and gallery face images must be the same prior to feature extraction [3]. As such, the given faces are normally resized to a common intermediate format (IF) prior to further processing[1] (*e.g.* low-resolution faces are upscaled while high-resolution faces are downscaled), and recognition systems are often tuned to work with that particular image size. The use of IF processing leads to three problems in mismatched resolution comparisons:

(i) For low-resolution images, upscaling does not introduce any new information, and can potentially intro-

*Acknowledgements: NICTA is funded by the Australian Government as represented by the *Department of Broadband, Communications and the Digital Economy* as well as the Australian Research Council through the *ICT Centre of Excellence* program.

[1] *cf.* intermediate frequency (IF) processing in a superheterodyne radio receiver.

duce artifacts or noise. Also, upscaled images are blurry (Fig. 1), which causes the extracted features to be very different than those obtained from the downscaled high-resolution faces, resulting in a significant drop in recognition accuracy [5]. Thus upscaling is generally not a good solution to the low-to-high resolution mismatch problem. It might be tempting to employ techniques such as super-resolution or hallucination [1], however super-resolution requires several images (which may not be available) in addition to precise alignment [9].

(ii) *Prima facie*, if upscaling is not a good solution, one may think that downscaling high-resolution images will solve the resolution mismatch issue. However, downscaling reduces the amount of information available, thereby reducing the recognition performance. Situations can arise where the given probe face image is larger than the IF image size (*e.g.* obtained through a telephoto lens). To allow maximum accuracy wherever possible, the recognition system should ideally be able to detect situations where a high-to-high resolution comparison is possible (*i.e.* with a larger IF) and when it should do a low-to-high resolution face comparisons (*i.e.* with a smaller IF). Typically, one IF processing chain alone is not sufficient to achieve this.

(iii) Resizing pre-supposes that the original sizes of the given images are an indicator of the underlying resolutions. This is often not the case in the poorly controlled image datasets encountered in practice. Thus a resolution detector is necessary to identify whether the underlying resolution of the probe image is high or low.

In this paper we present a novel method to handle resolution mismatches for the recently proposed Multi-Region Histograms (MRH) local-feature approach, which can be thought of as a hybrid between the HMM and GMM based systems [12]. Specifically, we propose: **(i)** the use of two IF sizes (small and large), with the small IF size targeted for reducing resolution mismatches caused by upscaling, and the large IF size targeted for high discrimination performance when little to no resolution mismatches are present; **(ii)** a dedicated resolution detector frontend to address situations where the actual resolution of given faces is unknown (*i.e.* where the size of given faces cannot be relied upon to determine the resolution); **(iii)** to employ the resolution detector, as part of a resolution mismatch compensation framework, to determine which of the two IF image sizes to use when comparing two face images with unknown resolutions.

We continue the paper as follows. In Section 2 we briefly describe the MRH-based face recognition approach. The proposed resolution mismatch compensation framework is described in Section 3. Section 4 presents experiments on the recent Labeled Faces in the Wild (LFW) dataset [8], which contains problematic face variations akin to those found in surveillance scenarios. The main findings are presented in Section 5.

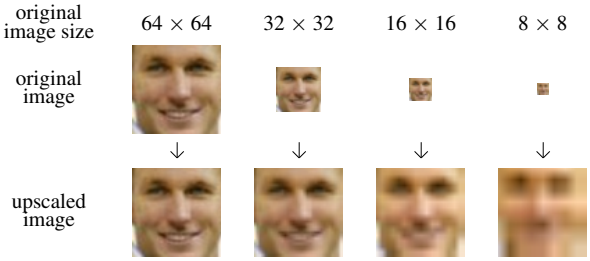

Figure 1. Original images of varying size upscaled to a size of 64×64 (via bilinear interpolation), resulting in images of fixed size but with varying underlying resolution.

2 Probabilistic Multi-Region Histograms

The MRH approach is motivated by the 'visual words' technique originally used in image categorisation [10]. Each face is divided into several fixed and adjacent regions, with each region comprising a relatively large part of the face. For region r a set of feature vectors is obtained, $F_r = \{\boldsymbol{f}_{r,i}\}_{i=1}^{N}$, which are in turn attained by dividing the region into small overlapping blocks (or patches) and extracting descriptive features from each block via 2D DCT decomposition [6]. Each block has a size of 8×8 pixels, which is the typical size used for DCT analysis. To account for varying contrast, each block is normalised to have zero mean and unit variance. Based on [12], coefficients from the top-left 4×4 sub-matrix of the 8×8 DCT coefficient matrix are used, excluding the 0-th coefficient (which has no information due to the normalisation).

For each vector $\boldsymbol{f}_{r,i}$ obtained from region r, a probabilistic histogram is computed:

$$\boldsymbol{h}_{r,i} = \left[\frac{w_1\, p_1\left(\boldsymbol{f}_{r,i}\right)}{\sum_{g=1}^{G} w_g p_g\left(\boldsymbol{f}_{r,i}\right)}, \cdots, \frac{w_G\, p_G\left(\boldsymbol{f}_{r,i}\right)}{\sum_{g=1}^{G} w_g p_g\left(\boldsymbol{f}_{r,i}\right)} \right]^T \quad (1)$$

where the g-th element in $\boldsymbol{h}_{r,i}$ is the posterior probability of $\boldsymbol{f}_{r,i}$ according to the g-th component of a visual dictionary model. As the visual dictionary is a mixture of Gaussians, the mean of each Gaussian can be thought of as a particular 'visual word'.

Once the histograms are computed for each feature vector from region r, an average histogram for the region is built:

$$\boldsymbol{h}_{r,\mathtt{avg}} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \boldsymbol{h}_{r,i} \quad (2)$$

The overlapping during feature extraction, as well as the loss of spatial relations within each region (due to averaging), results in robustness to translations of the face which are caused by imperfect face localisation. The DCT decomposition acts like a low-pass filter, with the information retained from each block being robust to small alterations (*e.g.* due to minor in-plane rotations).

The normalised distance between faces X and Y is calculated using:

$$d_{\mathtt{norm}}(X,Y) = \frac{d_{\mathtt{raw}}(X,Y)}{\frac{1}{2M}\sum_{i=1}^{M}\{d_{\mathtt{raw}}(X,C_i) + d_{\mathtt{raw}}(Y,C_i)\}} \quad (3)$$

where C_i is the i-th cohort face and M is the number of cohorts, while $d_{\mathtt{raw}}(\cdot,\cdot)$ is a L_1-norm based distance measure between histograms from R regions:

$$d_{\mathtt{raw}}(X,Y) = \frac{1}{R}\sum_{r=1}^{R}\left\|\boldsymbol{h}_{r,\mathtt{avg}}^{[X]} - \boldsymbol{h}_{r,\mathtt{avg}}^{[Y]}\right\|_1 \qquad (4)$$

Cohort faces are assumed to be reference faces that are known not to be of persons depicted in X or Y. The denominator in Eqn. (3) estimates how far away, on average, faces X and Y are from a randomly selected face. This typically results in Eqn. (3) being approximately 1 when X and Y represent faces from two different people, and less than 1 when X and Y represent two instances of the same person.

3 Proposed Compensation Framework

In order to handle resolution mismatches when the size of given face images cannot be relied upon as an indicator of the underlying resolution, it is necessary to analyse the content of the given images and determine whether the images can be downscaled to a more appropriate size.

In this work we use two IF image sizes, namely A and B. We define size A as 64×64 and size B as 32×32. It is important to note that due to the low-pass filtering effect of the DCT analysis, MRH-based recognition tuned for size A (where all given images are resized to size A) is able to handle images which have an underlying resolution ranging from 32×32 to 64×64, while MRH-based recognition tuned for size B is suited for 32×32 and lower resolutions (*i.e.* 16×16 and 8×8).

The detector uses two sets of reference faces: S_A and S_B. In each set the faces have a canonical size of 64×64 pixels, though in each set the underlying resolution is different. Set S_A contains faces which are downscaled versions of the underlying high resolution faces. In set S_B the underlying high resolution faces were first downscaled to 16×16, followed by upscaling to the canonical size (*i.e.* deliberate loss of information).

The detector co-opts the framework and processing used by the MRH approach, in order to exploit the sensitivity of local DCT features to resolution mismatches. In essence, the detector measures whether a given face is more similar to either low-resolution or high-resolution reference faces. The processing steps are:

1. The given face Q is rescaled to the canonical size (64×64), regardless of the original size of Q.
2. MRH analysis with 3×3 regions and 1024 visual words is performed (using parameter settings as in [12]).
3. The average distance of Q to faces in sets S_A and S_B is found:
$$d_{\mathtt{avg}}(Q,S_i) = |S_i|^{-1}\sum_{j=1}^{|S_i|} d_{\mathtt{raw}}\left(Q,S_{i,j}\right) \qquad (5)$$
where $i \in \{A,B\}$, $S_{i,j}$ is the j-th face of set S_i and $|S_i|$ is the number of faces in set S_i.
4. The smallest average distance, either $d_{\mathtt{avg}}(Q,S_A)$ or $d_{\mathtt{avg}}(Q,S_B)$, indicates whether MRH tuned for either size A or B, respectively, should be used for recognition.

Table 1. Classification performance of the proposed image resolution detector frontend. All given face images have one size (64×64) but the underlying resolution varies (8×8 to 64×64). Face images are classified as being suitable for MRH-based face recognition using either size A or B. MRH tuned for size A is suitable for images with an underlying resolution of 32×32 or higher, while MRH tuned for size B is more suited for lower resolutions.

Underlying Resolution	Size A	Size B
64×64	99.87 %	0.13 %
32×32	98.06 %	1.94 %
16×16	1.94 %	98.06 %
8×8	0.00 %	100.00 %

4 Experiments and Discussion

We used the Labeled Faces in the Wild (LFW) dataset which contains 13,233 face images (from 5749 unique persons) collected from the Internet [8]. The faces exhibit several compound problems such as misalignment and variations in pose, expression and illumination. In our experiments we extracted closely cropped faces[2] (to exclude the background) using a fixed bounding box placed in the same location in each LFW image.

In LFW experiment protocols the task is to classify a pair of previously unseen faces as either belonging to the same person or two different persons [8]. Performance is indicated by the mean of the accuracies from 10 folds of the 10 sets from view 2, in a leave-one-out cross-validation scheme (*i.e.* in each fold 9 sets are used for training and 1 set for testing, with each set having 300 same-person and 300 different-person pairs).

To study the effect of resolution mismatches, the first image in the each pair was rescaled to 64×64 while the second image was first rescaled to a size equal to or smaller than 64×64, followed by upscaling to the same size as the first image (*i.e.* deliberate loss of information, causing the image size to be uninformative as to the underlying resolution). The underlying resolution of the second image varied from 8×8 to 64×64.

In experiment 1 we evaluated the classification performance of the proposed resolution detector frontend. Reference faces for sets S_A and S_B were taken from the training set. Preliminary experiments indicated that using 32 faces for each reference set was sufficient. The second image in each pair from the test set was then classified as being suitable for MRH-based face recognition using either size A or B. Recall that an MRH-based face recognition system tuned for size A is suited for faces which have an underlying resolution of 32×32 or higher, while a corresponding system tuned for size B is more suited for lower resolutions. The results, shown in Table 1, indicate that the frontend detector is able to assign the most suitable size almost perfectly.

[2] Available from *http://itee.uq.edu.au/~conrad/lfwcrop/*

In experiment 2 we evaluated the performance of three MRH-based systems for classifying LFW image pairs subject to resolution mismatches. Systems A and B were tuned for size A and B, respectively, while the dynamic system applies the proposed compensation framework to switch between System A and B according to the classification result of the resolution detector.

Comparing the results of the two baseline systems (A and B) in Table 2 confirms that System A outperforms System B when matching images of similar underlying resolution (*i.e.* 64×64 and 32×32), but significantly underperforms System B when there is a considerable resolution mismatch (16×16 or lower). System B is able to achieve more rounded performance at the expense of reduced accuracy for the highest resolution (64×64).

The proposed dynamic system is able to retain the best aspect of System A (*i.e.* good accuracy at the highest resolution) with performance similar to System B at lower resolutions. Consequently, the dynamic system obtains the best overall performance.

We note that in three out of the four tested resolutions, the dynamic system slightly outperforms the best underlying system. Based on observations of the original LFW dataset, we conjecture that this outperformance is due to a subset of LFW images already having a low underlying resolution.

5 Conclusion

In this paper we have shown how comparing images with different underlying resolutions can lead to a significant drop in performance for a local feature based face recognition system, and proposed a compensation framework to improve overall performance (across several resolutions). The proposed framework relies on a novel resolution detector frontend which exploits the sensitivity of local features to resolution. The performance of this resolution detection and compensation framework was demonstrated on a resolution-modified Labeled Faces in the Wild [8] dataset using the Multi-Region Histogram based recognition system.

In our experiments, two systems (A and B) were tuned to different underlying resolutions. System A, tuned for higher underlying resolutions, was shown to outperform System B when comparing images of similar underlying resolution (64×64 and 32×32), while underperforming when comparing images of very different underlying resolution (16×16 and 8×8). The reverse was true for System B, tuned for lower resolutions. The proposed dynamic compensation framework was able to maximise performance by applying the system best tuned for any given pair of images based on their underlying resolutions. This potential to utilise the strengths of multiple face recognition systems clearly demonstrates the advantage of the compensation framework.

Table 2. Performance of three MRH-based systems for classifying LFW image pairs with resolution mismatches. All images had a fixed size of 64×64, but in each pair the second image had the underlying resolution varying from 8×8 to 64×64 (see Fig. 1). Systems A and B were tuned for size A and B, respectively, while the dynamic system switched between system A and B according to the classification result of the resolution detector.

Underlying Resolution	System A	System B	Dynamic System
64×64	74.25 %	70.28 %	74.35 %
32×32	70.36 %	69.99 %	70.47 %
16×16	59.35 %	68.08 %	67.62 %
8×8	53.13 %	59.40 %	59.90 %
Average	64.27 %	66.94 %	68.09 %

For a given pair of resolution-modified images from the LFW dataset, the proposed resolution detector was able to classify which face recognition system was the optimal one 99% of the time on average. This indicates nearly perfect face recognition system selection when used in the compensation framework.

References

[1] S. Baker and T. Kanade. Hallucinating faces. In *IEEE Int. Conf. Automatic Face and Gesture Recogition (AFGR)*, 2000.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, 1997.

[3] K. Bowyer. Face recognition technology: Security vs privacy. *IEEE Technology and Society Magazine*, 23(1):9–19, 2004.

[4] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Trans. Signal Processing*, 54(1):361–373, 2006.

[5] J. Choi, Y. Ro, and K. Plataniotis. Feature subspace determination in video-based mismatched face recognition. In *IEEE Int. Conf. Automatic Face and Gesture Recogition (AFGR)*, 2008.

[6] R. Gonzales and R. Woods. *Digital Image Processing*. Prentice Hall, 3rd edition, 2007.

[7] P. Hennings-Yeomans, S. Baker, and B. Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

[9] F. Lin, C. Fookes, V. Chandran, and S. Sridharan. Super-resolved faces for improved face recognition from surveillance video. In *Int. Conf. Biometrics (ICB), Lecture Notes in Computer Science (LNCS)*, volume 4642, pages 1–10, 2007.

[10] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conf. Computer Vision (ECCV), Part IV, Lecture Notes in Computer Science (LNCS)*, volume 3954, pages 490–503, 2006.

[11] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariethoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24(8):882–893, 2006.

[12] C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Int. Conf. Biometrics (ICB), Lecture Notes in Computer Science (LNCS)*, volume 5558, pages 199–208, 2009.