Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference

Conrad Sanderson[1,2] and Brian C. Lovell[1,2]

[1] NICTA, PO Box 6020, St Lucia, QLD 4067, Australia
[2] The University of Queensland, School of ITEE, QLD 4072, Australia

Abstract. We propose a scalable face matching algorithm capable of dealing with faces subject to several concurrent and uncontrolled factors, such as variations in pose, expression, illumination, resolution, as well as scale and misalignment problems. Each face is described in terms of multi-region probabilistic histograms of visual words, followed by a normalised distance calculation between the histograms of two faces. We also propose a fast histogram approximation method which dramatically reduces the computational burden with minimal impact on discrimination performance. Experiments on the "Labeled Faces in the Wild" dataset (unconstrained environments) as well as FERET (controlled variations) show that the proposed algorithm obtains performance on par with a more complex method and displays a clear advantage over predecessor systems. Furthermore, the use of multiple regions (as opposed to a single overall region) improves accuracy in most cases, especially when dealing with illumination changes and very low resolution images. The experiments also show that normalised distances can noticeably improve robustness by partially counteracting the effects of image variations.

1 Introduction

When dealing with images obtained in surveillance contexts (e.g. via CCTV), automatic identity inference based on faces is considerably more difficult than in well controlled conditions (e.g. immigration checkpoints). The difficulties arise due to several concurrent and uncontrolled factors: pose (this includes both in-plane and out-of-plane rotations), expression, illumination and resolution (due to variable distances to cameras). Furthermore, an automatic face locator (detector) must be used which can induce further problems. As there are no guarantees that the localisation is perfect, faces can be at the wrong scale and/or misaligned [1].

A surveillance system may have further constraints: only one gallery image per person, as well as real-time operation requirements in order to handle large volumes of people (e.g. peak hour at a railway station). In this context the computational complexity of an identity inference system is necessarily limited, suggesting that time-expensive approaches, such as the deduction of 3D shape from 2D images [2] (to compensate for pose variations), may not be applicable.

⋆ *Appears in: ICB 2009, LNCS 5558, pp. 199-208, 2009.*

In this work we describe a Multi-Region Histogram (MRH) based approach[3], with the aim of concurrently addressing the above-mentioned problems. The MRH approach is an evolution of a method presented in [3], which in turn was inspired by 'visual words' used in image categorisation [4]. The method presented here builds on [3] primarily through **(i)** multiple regions to increase discrimination performance without adversely affecting robustness, **(ii)** a histogram approximation method in order to dramatically speed up calculation with minimal impact on discrimination performance, and **(iii)** a distance normalisation method to improve robustness in uncontrolled image conditions.

We continue the paper as follows. Section 2 describes the proposed MRH approach in detail, along with the associated histogram approximation and distance normalisation methods. In Section 3 the MRH approach is briefly contrasted to related methods, taking scalability into account. Results from evaluations and comparisons on the Labeled Faces in the Wild (LFW) [5] and FERET [6] datasets are given in Section 4. The main findings and an outlook are presented in Section 5.

2 Multi-Region Histograms of Visual Words

Each face is divided into several fixed and adjacent regions, with each region comprising a relatively large part of the face (see Fig. 1). For region r a set of feature vectors is obtained, $X_r = \{\boldsymbol{x}_{r,1},\ \boldsymbol{x}_{r,2}, \cdots, \boldsymbol{x}_{r,N}\}$, which are in turn attained by dividing the region into small blocks (or patches) and extracting descriptive features from each block via 2D DCT [7] decomposition. Each block has a size of 8×8 pixels and overlaps neighbouring blocks by 75%. To account for varying contrast, each block is normalised to have zero mean and unit variance. Based on preliminary experiments we elected to retain 15 of the 64 DCT coefficients, by taking the top-left 4×4 submatrix of the 8×8 coefficient matrix and disregarding the first coefficient (as it carries no information due to the above normalisation).

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

$$\boldsymbol{h}_{r,i} = \left[\frac{w_1 p_1\left(\boldsymbol{x}_{r,i}\right)}{\sum_{g=1}^{G} w_g p_g\left(\boldsymbol{x}_{r,i}\right)},\ \frac{w_2 p_2\left(\boldsymbol{x}_{r,i}\right)}{\sum_{g=1}^{G} w_g p_g\left(\boldsymbol{x}_{r,i}\right)},\ \cdots,\ \frac{w_G p_G\left(\boldsymbol{x}_{r,i}\right)}{\sum_{g=1}^{G} w_g p_g\left(\boldsymbol{x}_{r,i}\right)} \right]^T \tag{1}$$

where the g-th element in $\boldsymbol{h}_{r,i}$ is the posterior probability of $\boldsymbol{x}_{r,i}$ according to the g-th component of a visual dictionary model. The visual dictionary model employed here is a convex mixture of gaussians [8], parameterised by $\lambda = \{w_g, \boldsymbol{\mu}_g, \mathbf{C}_g\}_{g=1}^{G}$, where G is the number of gaussians, while w_g, $\boldsymbol{\mu}_g$ and $\mathbf{C}_g$ are, respectively, the weight, mean vector and covariance matrix for gaussian g. 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:

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

[3] The source code for MRH can be obtained from *http://arma.sourceforge.net/mrh/*

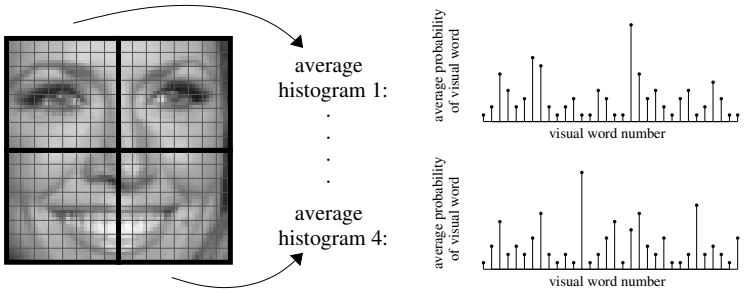


Fig. 1. Conceptual example of MRH face analysis using 2×2 regions, where each region is divided into small blocks. For each block descriptive features are placed into a vector. The posterior probability of the vector is then calculated using each gaussian in a visual dictionary, resulting in a histogram of probabilities. For each region the histograms of the underlying blocks are then averaged.

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 in-plane rotations, expression changes or smoothing due to upsampling from low resolution images). 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. We note that in the 1×1 region configuration (used in [3]) the overall topology of the face is effectively lost, while in configurations such as 3×3 it is largely retained (while still allowing for deformations in each region).

The visual dictionary is obtained by pooling a large number of feature vectors from training faces, followed by employing the Expectation Maximisation algorithm [8] to optimise the dictionary's parameters (i.e. λ).

2.1 Normalised Distance

Comparison of two faces can be accomplished by comparing their corresponding average histograms. Based on [9] we define an L_1-norm based distance measure between faces A and B:

$$d_{\texttt{raw}}(A,B) = \frac{1}{R}\sum\nolimits_{r=1}^{R} \left\| \boldsymbol{h}_{r,\texttt{avg}}^{[A]} - \boldsymbol{h}_{r,\texttt{avg}}^{[B]} \right\|_1 \tag{3}$$

where R is the number of regions. To reach a decision as to whether faces A and B come from the same person or from two different people, $d_{\texttt{raw}}(A,B)$ can be compared to a threshold. However, the optimal threshold might be dependent on the image conditions of face A and/or B, which are not known a-priori. Inspired by cohort normalisation [10], we propose a normalised distance in order to reduce the sensitivity of threshold selection:

$$d_{\texttt{normalised}}(A,B) = \frac{d_{\texttt{raw}}(A,B)}{\frac{1}{2}\left(\frac{1}{M}\sum_{i=1}^{M} d_{\texttt{raw}}(A,C_i) + \frac{1}{M}\sum_{i=1}^{M} d_{\texttt{raw}}(B,C_i)\right)} \tag{4}$$

where C_i is the i-th cohort face and M is the number of cohorts. In the above equation cohort faces are assumed to be reference faces that are known not to

be of persons depicted in A or B. As such, the terms $\frac{1}{M}\sum_{i=1}^{M} d_{\texttt{raw}}(A, C_i)$ and $\frac{1}{M}\sum_{i=1}^{M} d_{\texttt{raw}}(B, C_i)$ estimate how far away, on average, faces A and B are from the face of an impostor. This typically results in Eqn. (4) being approximately 1 when A and B represent faces from two different people, and less than 1 when A and B represent two instances of the same person. If the conditions of given images cause their raw distance to increase, the average raw distances to the cohorts will also increase. As such, the division in Eqn. (4) attempts to cancel out the effect of varying image conditions.

2.2 Fast Histogram Approximation

As will be shown in Section 4, the size of the visual dictionary needs to be relatively large in order to obtain good performance. Typically about 1000 components (gaussians) are required, which results in the calculation of histograms via Eqn. (1) to be time consuming. Based on empirical observations that for each vector only a subset of the gaussians is dominant, we propose a dedicated algorithm that adaptively calculates only a part of the histogram. The algorithm is comprised of two parts, with the first part done during training.

In the first part, the gaussians from the visual dictionary model are placed into K clusters via the k-means algorithm [8]. Euclidean distance between the means of the gaussians is used in determining cluster memberships. For each cluster, the closest member to the cluster mean is labelled as a *principal gaussian*, while the remaining members are labelled as *support gaussians*.

For a feature vector $\boldsymbol{x}$ an approximate histogram is then built as follows. Each of the K principal gaussians is evaluated. The clusters are then ranked according to the likelihood obtained by each cluster's principal gaussian (highest likelihood at the top). Additional likelihoods are produced cluster by cluster, with the production of likelihoods stopped as soon as the total number of gaussians used (principal and support) exceeds a threshold. The histogram is then constructed as per Eqn. (1), with the likelihoods of the omitted gaussians set to zero.

3 Related Methods and Scalability

The use of probabilistic histograms in MRH differs to the histograms used in [4] (for image retrieval/categorisation purposes), where a Vector Quantiser (VQ) based strategy is typically used. In the VQ strategy each vector is forcefully assigned to the closest matching visual word, instead of the probabilistic assignment done here.

For the purposes of face classification, MRH is related to, but distinct from, the following approaches: Partial Shape Collapse (PSC) [11], pseudo-2D Hidden Markov Models (HMMs) [12, 13] and Probabilistic Local PCA (PLPCA) [14].

MRH is also somewhat related to the recently proposed and more complex Randomised Binary Trees (RBT) method [15], aimed for more general object classification. While both MRH and RBT use image patches for analysis, RBT also uses: (i) quantised differences, via 'extremely-randomised trees',

between corresponding patches, (ii) a cross-correlation based search to determine patch correspondence, and (iii) an SVM classifier [8] for final classification.

The differences between MRH and PSC include: (i) the use of fixed regions for all persons instead of manually marked regions for each person, (ii) each region is modelled as a histogram rather than being directly described by a Gaussian Mixture Model (GMM), leading to (iii) MRH using only one GMM (the visual dictionary), common to all regions and all persons, instead of multiple GMMs per person in PSC. The use of only one GMM directly leads to much better scalability, as the number of gaussians requiring evaluation for a given probe face is fixed, rather than growing with the size of the face gallery. In the latter case the computational burden can quickly become prohibitive [3, 10].

The MRH approach has similar advantages over PLPCA and HMM in terms of scalability and histogram based description. However, there are additional differences. In PLPCA each region is analysed via PCA instead of being split into small blocks. While the probabilistic treatment in PLPCA affords some robustness to translations, the use of relatively large face areas is likely to have negative impact on performance when dealing with other image transformations (e.g. rotations and scale changes). In HMM approaches the region boundaries are in effect found via an automatic alignment procedure (according to the model of each person the face is evaluated against) while in the MRH approach the regions are fixed, allowing straightforward parallel processing.

4 Experiments

The experiments were done on two datasets: LFW [5], and subsets of FERET [6]. We will show results of LFW first, where the number of face variations (as well as their severity) is uncontrolled, followed by a more detailed study on FERET, where each variation (e.g. pose) is studied separately.

The recent LFW dataset contains 13,233 face images which have several compound problems – e.g. in-plane rotations, non-frontal poses, low resolution, non-frontal illumination, varying expressions as well as imperfect localisation, resulting in scale and/or translation issues. The images were obtained by trawling the Internet followed by face centering, scaling and cropping based on bounding boxes provided by an automatic face locator. The original bounding boxes were expanded to include context. In our experiments we extracted closely cropped faces using a fixed bounding box placed in the same location in each LFW image[4]. The extracted faces were size normalised to 64×64 pixels, with an average distance between the eyes of 32 pixels. Examples are shown in Fig. 2.

LFW experiments follow a prescribed protocol [5], where the task is to classify a pair of previously unseen faces as either belonging to the same person (matched pair) or two different persons (mismatched pair). The protocol specifies

[4] The upper-left and lower-right corners of the bounding box were: (83,92) and (166,175), respectively. Bounding box location was determined first via centering then shifting upwards to fit the eyes and mouth of 40 randomly selected LFW faces.

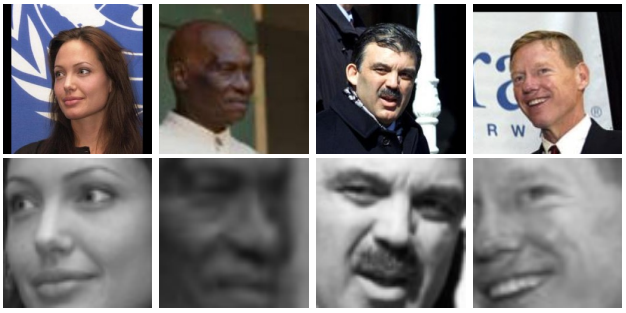


Fig. 2. Examples from the LFW dataset: **(i)** master images (resulting from centering, scaling and cropping based on bounding boxes provided by an automatic face locator); **(ii)** processed versions used in experiments, extracted using a fixed bounding box placed on the master images. The faces typically have at least one of the following issues: in-plane rotations, non-frontal poses, low resolution, non-frontal illumination, varying expressions as well as imperfect localisation, resulting in scale and/or translation issues.

two views of the dataset: *view 1*, aimed at algorithm development & model selection, and *view 2*, aimed at final performance reporting (to be used sparingly). In view 1 the images are split into two sets: the training set (1100 matched and 1100 mismatched pairs) and the testing set (500 matched and 500 mismatched pairs). The training set is used for constructing the visual dictionary as well as selecting the decision threshold. The threshold was optimised to obtain the highest *average accuracy* (averaged over the classification accuracies for matched and mismatched pairs). In view 2 the images are split into 10 sets, each with 300 matched & 300 mismatched pairs. Performance is reported using the mean and standard error of the average accuracies from 10 folds of the sets, in a leave-one-out cross-validation scheme (i.e. in each fold 9 sets are used training and 1 set for testing). The standard error is useful for assessing the significance of performance differences across algorithms [5].

For consistency, experiments on FERET were designed to follow a similar pair classification strategy, albeit with manually found eye locations. The 'fa' and 'fb' subsets (frontal images) were used for training – constructing the visual dictionary as well as selecting the decision threshold. Using persons which had images in both subsets, there were 628 matched and 628 randomly assigned mismatched pairs. The 'b' subsets were used for testing, which contain controlled pose, expression and illumination variations for 200 unique persons. For each image condition there were 200 matched and 4000 mismatched pairs, with the latter obtained by randomly assigning 20 persons to each of the 200 available persons. Image transformations were applied separately to the frontal source images ('ba' series), obtaining the following versions: in-plane rotated (20°), scaled (bounding box expanded by 20% in x and y directions, resulting in shrunk faces), translated (shifted in x and y directions by 6 pixels, or 20% of the distance between the eyes), upsampled from a low resolution version (with the low resolution version obtained by shrinking the original image to 30% of its size, resulting in an average eye distance of ∼10 pixels). Example images are shown in Fig. 3.

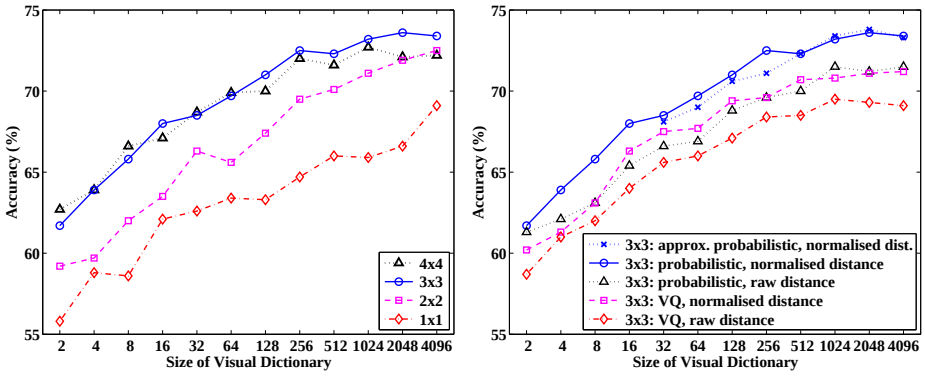
Fig. 3. *Top row:* examples of cropped images from FERET (neutral, followed by expression, illumination and pose change). *Bottom row:* transformed and cropped versions of the neutral source image (in-plane rotation, scale change, translation and upsampled low-res version).

In experiment 1 we studied the effect of increasing the size of the visual dictionary (from 2 to 4096 components) and number of regions (from 1×1 to 4×4) on the LFW dataset. As these variations constitute model selection, view 1 was used. The system used probabilistic histograms and normalised distances. Based on preliminary experiments, 32 randomly chosen cohort faces from the training set were used for the distance normalisation. The results, in Fig. 4(i), suggest that performance steadily increases up to about 1024 components, beyond which performance changes are mostly minor. Dramatic improvements are obtained by increasing the number of regions from 1×1 to 3×3. Using more regions (i.e. 4×4) shows no appreciable further performance gains.

In experiment 2 we fixed the number of regions at 3×3 and varied the size of the visual dictionary. The performance of systems using exact probabilistic histograms, approximate probabilistic and VQ based was compared. We also evaluated the performance of raw and normalised distances on both probabilistic and VQ based systems. Based on preliminary experiments, approximate histograms used $K=\frac{G}{10}$ clusters and a maximum of $\frac{G}{4}$ gaussians, where G is the size of the visual dictionary. The results, in Fig. 4(ii), point to the distance normalisation being helpful, with a consistent advantage of about 2 percentage points over raw distances (e.g. 72% vs 70%). The results further suggest that probabilistic histograms outperform VQ based histograms, also with an advantage of about 2 points. Finally, the performance of the computationally less expensive approximate probabilistic histograms is on par with exact probabilistic histograms.

In experiment 3 we used view 2 of LFW, allowing comparison with previously published as well as future results. Several configurations of MRH were evaluated as well as a baseline PCA system. Based on preliminary experiments, the baseline PCA based system used the euclidean distance as its raw distance and 61 eigenfaces (eigenfaces 4 to 64 of the training images, following the recommendation in [16] to skip the first three eigenfaces). The results, presented in Table 1, indicate that the performance of MRH based systems is consistent with experiments 1 and 2. Furthermore, the probabilistic 3×3 MRH method is on par with the more complex RBT method. The performance of PCA considerably lags behind all other approaches.

In experiment 4 images from FERET were used. The performances of probabilistic MRH with 3×3 and 1×1 configurations, as well as the baseline PCA based system, were compared. Both raw and normalised distances were evalu-



(i) **(ii)**

Fig. 4. Accuracy rate for increasing size of the visual dictionary, on *view 1* of LFW (compound face variations). **(i)**: MRH (probabilistic, normalised distance), with the number of regions varying from 1×1 to 4×4. **(ii)**: 3×3 MRH, with either probabilistic or VQ based histogram generation, as well with and without distance normalisation.

Method	**Mean accuracy**	**Standard error**
3×3 MRH (approx probabilistic, normalised distance)	72.35	0.54
3×3 MRH (probabilistic, normalised distance)	**72.95**	0.55
3×3 MRH (probabilistic, raw distance)	70.38	0.48
3×3 MRH (VQ, normalised distance)	69.35	0.72
3×3 MRH (VQ, raw distance)	68.38	0.61
1×1 MRH (probabilistic, normalised distance)	67.85	0.42
PCA (normalised distance)	59.82	0.68
PCA (raw distance)	57.23	0.68
Randomised Binary Trees (RBT)	72.45	0.40

Table 1. Results on *view 2* of LFW. Results for RBT obtained from http://vis-www.cs.umass.edu/lfw (accessed 2008-09-01), using the method published in [15]. MRH approaches used a 1024 component visual dictionary.

ated. For testing, each image condition was evaluated separately. Moreover, for each image pair to be classified, the first image was always from the 'ba' series (normal frontal image). The results, presented in Fig. 5, indicate that increasing the number of regions from 1×1 to 3×3 improves accuracy in most cases, especially when dealing with illumination changes and low resolution images. The notable exceptions are faces with pose changes and in-plane rotations. We conjecture that the histogram for each region (3×3 case) is highly specific and that it has simply altered too much due to the pose change; in the 1×1 case, the single overall histogram is more general and parts of it are likely to be describing face sections which have changed relatively little. For the in-plane rotation, we conjecture that the performance drop is at least partially due to face components (e.g. eyes) moving between regions, causing a mismatch between the corresponding histograms of two faces; in the 1×1 case there is only one histogram, hence the movement has a reduced effect.

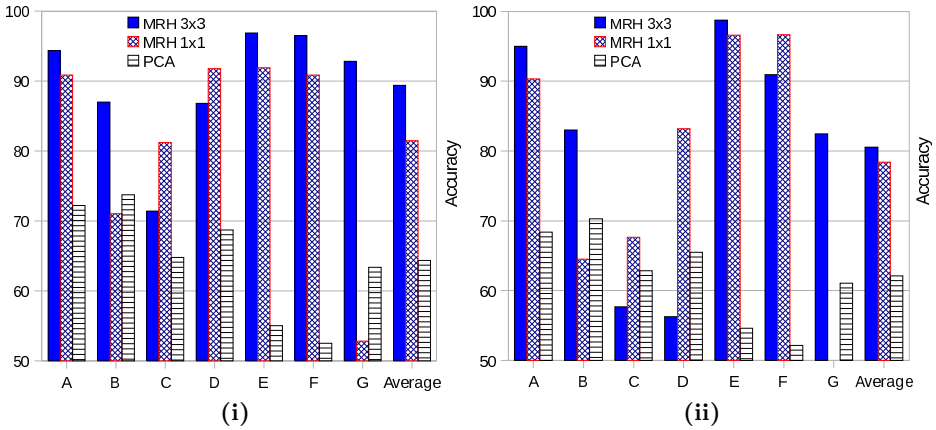


Fig. 5. Performance on FERET (separate face variations), using: **(i)** normalised distances, **(ii)** raw distances. A: expression change, B: illumination change, C: pose change, D: in-plane rotation, E: scale change, F: translation (shift), G: upsampled low resolution image.

The use of normalised distances improved the average performance of all approaches. This is especially noticeable for MRH when dealing with pose changes and in-plane rotations. In all cases the 3×3 MRH system considerably outperformed the baseline PCA system, most notably for faces subject to scale changes and translations.

5 Main Findings and Outlook

In this paper we proposed a face matching algorithm that describes each face in terms of multi-region probabilistic histograms of visual words, followed by a normalised distance calculation between the corresponding histograms of two faces. We have also proposed a fast histogram approximation method which dramatically reduces the computational burden with minimal impact on discrimination performance. The matching algorithm was targeted to be scalable and deal with faces subject to several concurrent and uncontrolled factors, such as variations in pose, expression, illumination, as well as misalignment and resolution issues. These factors are consistent with face images obtained in surveillance contexts.

Experiments on the recent and difficult LFW dataset (unconstrained environments) show that the proposed algorithm obtains performance on par with the recently proposed and more complex Randomised Binary Trees method [15]. Further experiments on FERET (controlled variations) indicate that the use of multiple adjacent histograms (as opposed to a single overall histogram) on one hand reduces robustness specific to in-plane rotations and pose changes, while on the other hand results in better average performance. The experiments also show that use of normalised distances can considerably improve the robustness of both multiple- and single-histogram systems.

The robustness differences between multiple- and single-histogram systems suggest that combining the two systems (e.g. by a linear combination of distances) could be beneficial. Lastly, we note that the MRH approach is easily amenable to parallelisation: a multi-CPU machine can process regions concurrently, thereby providing a significant speed-up.

Acknowledgements

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

References

1. Rodriguez, Y., Cardinaux, F., Bengio, S., Mariethoz, J.: Measuring the performance of face localization systems. Image and Vision Comput. **24** (2006) 882–893
2. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Analysis and Machine Intelligence **25**(9) (2003) 1063–1074
3. Sanderson, C., Shan, T., Lovell, B.C.: Towards pose-invariant 2D face classification for surveillance. In: Analysis and Modeling of Faces and Gestures (AMFG), Lecture Notes in Computer Science (LNCS). Volume 4778. (2007) 276–289
4. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: European Conference on Computer Vision (ECCV), Part IV, Lecture Notes in Computer Science (LNCS). Volume 3954. (2006) 490–503
5. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007
6. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Analysis and Machine Intelligence **22**(10) (2000) 1090–1104
7. Gonzales, R., Woods, R.: Digital Image Processing. 3 edn. Prentice Hall (2007)
8. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
9. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision **45**(2) (2001) 83–105
10. Sanderson, C.: Biometric Person Recognition — Face, Speech and Fusion. VDM Verlag (2008)
11. Lucey, S., Chen, T.: A GMM parts based face representation for improved verification through relevance adaptation. In: Proc. Computer Vision and Pattern Recognition (CVPR). Volume 2. (2004) 855–861
12. Nefian, A., Hayes, M.: Face recognition using an embedded HMM. In: Proc. Audio Video-based Biometric Person Authentication (AVBPA). (1999) 19–24
13. Cardinaux, F., Sanderson, C., Bengio, S.: User authentication via adapted statistical models of face images. IEEE Trans. Signal Processing **54**(1) (2006) 361–373
14. Martínez, A.M.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. Pattern Analysis and Machine Intelligence **24**(6) (2002) 748–763
15. Nowak, E., Jurie, F.: Learning visual similarity measures for comparing never seen objects. In: Proc. Computer Vision and Pattern Recognition (CVPR). (2007) 1–8
16. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Analysis and Machine Intelligence **19**(7) (1997) 711–720