Combined Learning of Salient Local Descriptors and Distance Metrics for Image Set Face Verification

Conrad Sanderson, Mehrtash T. Harandi, Yongkang Wong, Brian C. Lovell

NICTA, PO Box 6020, St Lucia, QLD 4067, Australia
University of Queensland, School of ITEE, QLD 4072, Australia

Abstract

In contrast to comparing faces via single exemplars, matching sets of face images increases robustness and discrimination performance. Recent image set matching approaches typically measure similarities between subspaces or manifolds, while representing faces in a rigid and holistic manner. Such representations are easily affected by variations in terms of alignment, illumination, pose and expression. While local feature based representations are considerably more robust to such variations, they have received little attention within the image set matching area. We propose a novel image set matching technique, comprised of three aspects: ***(i)*** *robust descriptors of face regions based on local features, partly inspired by the hierarchy in the human visual system,* ***(ii)*** *use of several subspace and exemplar metrics to compare corresponding face regions,* ***(iii)*** *jointly learning which regions are the most discriminative while finding the optimal mixing weights for combining metrics. Experiments on LFW, PIE and MOBIO face datasets show that the proposed algorithm obtains considerably better performance than several recent state-of-the-art techniques, such as Local Principal Angle and the Kernel Affine Hull Method.*

1. Introduction

A recent trend in image set matching considers image sets as linear subspaces, with the similarity between the sets derived from the similarity between the subspaces [5, 13, 35, 36]. In almost all subspace based approaches, faces are represented in a rigid and holistic manner, where each face is represented by one feature vector that describes the entire face. Such a representation implicitly embeds rigid spatial constraints between face components [4].

While subspaces are thought of being capable of accommodating the effects of various image variations[1], the magnitude and compounding effect of variations (such as illumination, pose and expression changes) might overwhelm even the most sophisticated subspace modelling technique. The relatively poor performance of linear models in such challenging recognition tasks appears to have roots in the non-linear nature of typical image manifolds [19, 33, 35], with much effort directed towards handling the non-linearities (eg. via kernel extensions [13, 35] and data clustering [9, 11, 33]).

In contrast to rigid face representations, a face can also be represented by a set of local features. This set can then be processed by a classifier that explicitly allows relaxed spatial constraints between face parts. Such a combination allows for some movement and/or deformations of the face components [4, 16, 27], which in turn leads to a degree of inherent robustness to expression and pose changes [16, 27] as well as misalignment [4]. Examples of such systems include Elastic Graph Matching [34], pseudo-2D hidden Markov models [4], and "bag of words" approaches [28].

Several studies in the domain of single-image to single-image matching have shown that non-linear structures can be effectively avoided by local representations. More precisely, while the structures that describe holistic features tend to be non-linear and complex, linear structures are good/sufficient tools to approximate local features [18, 24]. As such, rather than using holistic face representations and relying on a model to handle the resulting non-linear variations, it might be more appropriate to develop an image set matching technique based on local representations, while allowing relaxed spatial constraints.

We propose an approach for image set face verification that uses a multitude of local representations and distance metrics, and employs a learning algorithm to determine which subset of descriptors and their associated metrics is the most useful for discrimination. More specifically, from each image two types of robust local descriptors are obtained: region descriptors and compound region descriptors. The compound descriptors are inspired by the hierarchical architecture of the human visual system, where the receptive fields of neurons tend to get larger in order to deal with increasingly complex stimuli [29]. The descriptors from corresponding regions in two face sets are pooled and then compared via several distance metrics (instead of relying on only one), resulting in a high-dimensional similarity vector. As such, the image set verification problem is converted to a binary problem on similarity features.

By learning to separate similarity vectors representing matched sets (ie. sets of the same person) and mismatched sets (ie. sets of two persons), we are in effect jointly determining which regions are the most discriminative while finding the optimal mixing weights for combining metrics. Fig. 1 shows a conceptual overview of the approach.

[1] For example, a linear subspace can be used for photometric invariance, under the conditions of no shadowing and Lambertian reflectance [1].

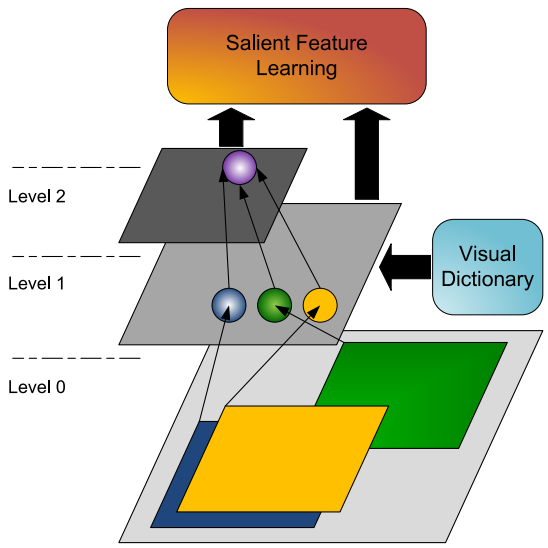


Figure 1. Partly inspired by the human visual system [29], the proposed approach has a hierarchical structure. Level 0 is the image plane. Level 1 contains descriptors for regions within the face, with the regions having arbitrary sizes and locations. Each descriptor is a probabilistic histogram, obtained using a dictionary of visual words. Level 2 contains compound descriptors, generated by aggregating the descriptors from Level 1. The descriptors from Level 1 and 2 are fed to a learning mechanism which determines which subset of descriptors is the most useful for face verification.

We continue the paper as follows. The feature extraction process is described in Section 2. The details of the learning approach are given in Section 3. Comparative evaluations of the proposed method against other image set matching techniques are given in Section 4. The main findings and possible future directions are covered in Section 5.

2. Hierarchical Feature Extraction

As shown in Fig. 1, the feature extraction is hierarchical in nature, with 3 levels. The lowest level (level 0), is the image plane. The details for the feature extraction at levels 1 and 2 are given in Sections 2.1 and 2.2, respectively.

2.1. Level 1

Each descriptor in level 1 corresponds to a relatively large region in the image plane. The descriptor for region size of $p \times p$, at an arbitrary location, is constructed as follows. In a similar manner to [28], the region is split into small overlapping blocks, with each block having a size of 8×8. For each block a histogram of probabilities is calculated, where each entry in the histogram reflects the similarity of the block to a pre-defined 'visual word'. Each region is represented as the average of all the histograms obtained for the region's blocks. The procedure is elucidated below.

Each block is represented by a low-dimensional texture descriptor. For each texture descriptor $\boldsymbol{x}_{r,i}$ obtained from a block in region r, a probabilistic histogram is computed:

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

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 [3], parameterised by $\lambda = \{w_g, \boldsymbol{\mu}_g, \mathbf{C}_g\}_{g=1}^{N_G}$, where N_G is the number of Gaussians, while w_g, $\boldsymbol{\mu}_g$ and $\mathbf{C}_g$ are the weight, mean vector and covariance matrix for Gaussian g, respectively. The mean of each Gaussian can be thought of as a particular 'visual word'. The visual dictionary is obtained by pooling a large number of texture descriptors from training images, followed by employing the Expectation Maximisation algorithm [3] to find the dictionary's parameters (*i.e.*, λ).

In this work we use local texture descriptors based on DCT analysis with illumination normalisation [28]. However, it is possible to use other texture descriptors, eg., based on Gabor wavelets [20] or Local Binary Patterns [2].

Once the histograms are computed for each feature vector from region r, an average histogram for the region is built: $\boldsymbol{h}_{r,\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}_{r,i}$. Due to the averaging operation, in each region there is a loss of spatial relations between face parts. As such, each region is in effect described by an orderless collection of local features ('bag-of-words'). The loss of spatial relations allows for a degree of misalignment, pose variations and expression changes [4, 16, 27, 28].

2.2. Level 2

In the human visual system, the receptive fields of neurons tend to get larger in order to deal with increasingly complex stimuli [29]. The responses of complex cells can be pooled from the responses of adjacent simple cells using 'max' or 'sum' operations [25, 29]. In a similar manner, we use three configurations for combining the descriptors from level 1, using the 'sum' operation.

The three configurations are shown in Fig. 2. The first configuration is in effect a horizontal shape. The compound descriptor in this case is a summation of three regions, *i.e.* simple cells, where the centers of the two outer regions are located at $(-(p-d), 0)$ and $(p-d, 0)$ relative to the center of the middle region, where $p \times p$ is the region size. The second configuration is similar to the first, except a vertical shape is used. We conjecture the first configuration can be useful for capturing horizontally elongated structures such as the mouth, eyes and eyebrows, while the second configuration can be useful for representing vertically elongated shapes, such as the nose.

The third configuration is a combination of the previous two shapes and forms a cross shape. We believe it can be useful for capturing a degree of correlations between the appearance of various face parts. For example, the shape of the nose might be related to the shape of the mouth.

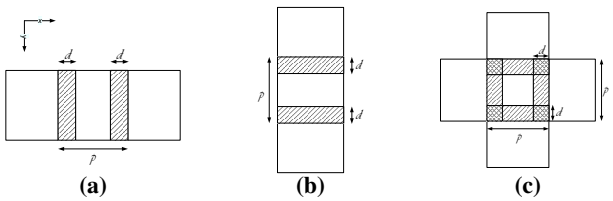


Figure 2. Compound descriptors are generated by computing the sum over the descriptors from level 1. We use three configurations: **(a)** for representing horizontally elongated shapes like the eyes and mouth; **(b)** for representing vertical elongated shapes such as the nose; **(c)** a mixture of (a) and (b), for capturing a degree of correlations between shapes such as the nose and mouth.

3. Determining Salient Descriptors

An image set face verification system needs to determine whether two sets, $\mathbb{A}$ and $\mathbb{B}$, represent the same person. In general this is accomplished by comparing the similarity between the two sets to a predefined threshold τ.

We assume that the image set $\mathbb{A}$ is comprised of l images. Each image i is represented by ν descriptors (histograms from level 1 and 2), $\boldsymbol{h}_1^{[i]}, \boldsymbol{h}_2^{[i]}, \cdots, \boldsymbol{h}_\nu^{[i]}$, with each descriptor covering a particular region. We define a *local mode* as a matrix which contains all descriptors for region j from the l images:

$$\boldsymbol{L}_j^{\mathbb{A}} = \left[\boldsymbol{h}_j^{[1]} \mid \boldsymbol{h}_j^{[2]} \mid \cdots \mid \boldsymbol{h}_j^{[l]}\right],\ j = 1, 2, \cdots, \nu \tag{2}$$

To compare two corresponding local modes from sets $\mathbb{A}$ and $\mathbb{B}$, ie., $\boldsymbol{L}_j^{\mathbb{A}}$ and $\boldsymbol{L}_j^{\mathbb{B}}$, instead of relying on only one similarity measure, we propose to use k similarity measures: $d_1\left(\boldsymbol{L}_j^{\mathbb{A}}, \boldsymbol{L}_j^{\mathbb{B}}\right)$, $d_2\left(\boldsymbol{L}_j^{\mathbb{A}}, \boldsymbol{L}_j^{\mathbb{B}}\right)$, $\cdots$, $d_k\left(\boldsymbol{L}_j^{\mathbb{A}}, \boldsymbol{L}_j^{\mathbb{B}}\right)$. We define the overall *similarity vector* between sets $\mathbb{A}$ and $\mathbb{B}$ as containing k similarity measures for each local mode, resulting in a $k\nu$-dimensional vector:

$$\boldsymbol{S}(\mathbb{A}, \mathbb{B}) = \begin{bmatrix} d_1(\boldsymbol{L}_1^{\mathbb{A}}, \boldsymbol{L}_1^{\mathbb{B}}) \\ \cdots \\ d_k(\boldsymbol{L}_1^{\mathbb{A}}, \boldsymbol{L}_1^{\mathbb{B}}) \\ d_1(\boldsymbol{L}_2^{\mathbb{A}}, \boldsymbol{L}_2^{\mathbb{B}}) \\ \cdots \\ d_k(\boldsymbol{L}_2^{\mathbb{A}}, \boldsymbol{L}_2^{\mathbb{B}}) \\ \cdots \\ d_1(\boldsymbol{L}_\nu^{\mathbb{A}}, \boldsymbol{L}_\nu^{\mathbb{B}}) \\ \cdots \\ d_k(\boldsymbol{L}_\nu^{\mathbb{A}}, \boldsymbol{L}_\nu^{\mathbb{B}}) \end{bmatrix} \tag{3}$$

The image set verification problem is hence converted to a binary classification problem involving similarity vectors. Figure 3 provides a graphical interpretation.

We use two families of similarity measures: subspace based, and exemplar based. For the subspace based measures, we employ the Grassmannian geodesic distance (arc-length) and Binet-Cauchy distance [12]. For the exemplar based measures, we use Hausdorff and Modified Hausdorff distances [7]. The two families are elucidated below.

For the subspace based measures, each local mode $\boldsymbol{L}_i^{\mathbb{A}}$ is modelled by a linear subspace. A common similarity measure between subspaces is the concept of principal angles [36]. If $\boldsymbol{O}_1 \in \mathbb{R}^{d \times n_1}$ and $\boldsymbol{O}_2 \in \mathbb{R}^{d \times n_2}$ are two linear subspaces in $\mathbb{R}^d$ with minimum rank $r = \min(rank(\boldsymbol{O}_1, \boldsymbol{O}_2))$, then there are exactly r uniquely defined principal angles between $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$:

$$\cos(\theta_i) = \max_{\boldsymbol{x}_i \in \boldsymbol{O}_1,\ \boldsymbol{y}_j \in \boldsymbol{O}_2} \boldsymbol{x}_i^T \boldsymbol{y}_j \tag{4}$$

subject to $\boldsymbol{x}_i^T \boldsymbol{x}_i = \boldsymbol{y}_i^T \boldsymbol{y}_i = 1, \boldsymbol{x}_i^T \boldsymbol{x}_j = \boldsymbol{y}_i^T \boldsymbol{y}_j = 0, i \neq j$. A straightforward method for computing the principal angles is based on Singular Value Decomposition. More specifically, the cosines of the principal angles are the singular values of $\boldsymbol{O}_1^T \boldsymbol{O}_2$:

$$\boldsymbol{O}_1^T \boldsymbol{O}_2 = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{V}^T \tag{5}$$

where the singular values are the diagonal entries of $\boldsymbol{\Lambda}$.

Based on the above principal angles, we use two similarity measures: Grassmannian geodesic distance and Binet-Cauchy distance, defined respectively as [12]:

$$d_{\mathrm{G}}(\boldsymbol{O}_1, \boldsymbol{O}_2) = \sum\nolimits_i \theta_i^2 \tag{6}$$

$$d_{\mathrm{BC}}(\boldsymbol{O}_1, \boldsymbol{O}_2) = \sqrt{1 - \prod\nolimits_i \cos^2(\theta_i)} \tag{7}$$

For the exemplar based measures, local modes are compared using Hausdorff and Modified Hausdorff distances [7]. Given two corresponding local modes $\boldsymbol{L}_i^{\mathbb{A}}$ and $\boldsymbol{L}_i^{\mathbb{B}}$, the Hausdorff distance (HD) is defined as:

$$d_{\mathrm{HD}}\left(\boldsymbol{L}_i^{\mathbb{A}}, \boldsymbol{L}_i^{\mathbb{B}}\right) = \max\left(\max_{a \in \mathbb{A}} \min_{b \in \mathbb{B}} \|a - b\|, \max_{b \in \mathbb{B}} \min_{a \in \mathbb{A}} \|a - b\|\right) \tag{8}$$

Intuitively, if the Hausdorff distance is d, then every point of $\mathbb{A}$ must be within a distance d of some point $\mathbb{B}$ and vice versa. For image processing applications, Dubuisson *et al.* [7] proposed the modified Hausdorff distance (MHD), which is more robust against outliers:

$$d_{\mathrm{MHD}}(\boldsymbol{L}_i^{\mathbb{A}}, \boldsymbol{L}_i^{\mathbb{B}}) = \max\left(d_{\mathrm{M}}(\boldsymbol{L}_i^{\mathbb{A}}, \boldsymbol{L}_i^{\mathbb{B}}), d_{\mathrm{M}}(\boldsymbol{L}_i^{\mathbb{B}}, \boldsymbol{L}_i^{\mathbb{A}})\right) \tag{9}$$

where $d_{\mathrm{M}}(\boldsymbol{L}_i^{\mathbb{A}}, \boldsymbol{L}_i^{\mathbb{B}}) = \frac{1}{|\mathbb{A}|} \sum_{a \in \mathbb{A}} \min_{b \in \mathbb{B}} \|a - b\|$, with $|\mathbb{A}|$ denoting the cardinality of set $\mathbb{A}$.

Due to the dense nature of the feature extraction process, a hefty and redundant representation is available for any image, leading to a very high dimensional similarity vector. A further contributing factor to the high dimensionality is the use of four distance metrics per local mode. As such, instead of blindly feeding the similarity vectors to a standard learning mechanism such as a Support Vector Machine [3], we have elected to use an adapted version of the AdaBoost algorithm [31], which is more suitable for dealing with such high dimensional problems.

In the adapted AdaBoost, each weak learner works for a single feature each time. As a result after Q rounds of boosting, Q features are selected. The adapted version hence has a considerably lower computational complexity than the original version [10]: in a D-dimensional problem, Q comparisons are required instead of $Q \times D$.

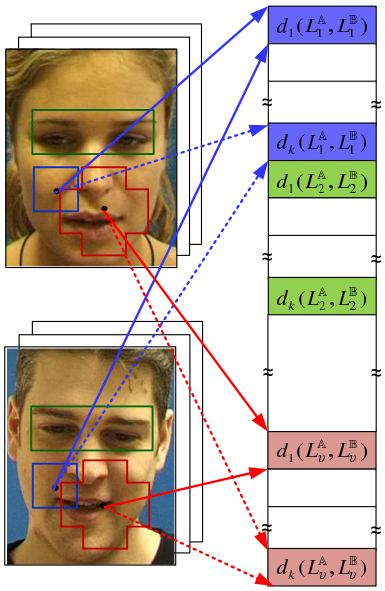


Figure 3. Converting the image set verification problem to a binary problem on similarity features. Each region in a single image for a given person is described by an average histogram of visual words. The corresponding histograms for a particular region across several images form a local mode. The corresponding local modes from two people are compared using several distance metrics. All the resulting distances for all modes are placed into a similarity vector.

4. Experiments

In this section we first provide an overview of the image datasets used in the experiments (Section 4.1), followed by a comparative performance evaluation against several benchmark and recent state-of-the-art methods (Section 4.2).

4.1. Image Datasets

We employed 3 datasets: Labeled Faces in the Wild (LFW) [17], CMU PIE [30] and MOBIO [22]. The datasets contain various face orientations, expressions, illumination situations and occlusions. A verification setup similar to the LFW protocol [17] is used, where the task is to classify a pair of previously unseen image sets as either belonging to the same person (matched pair) or two different persons (mismatched pair). In all experiments the images are split into three groups: **(i)** training, **(ii)** development, **(iii)** evaluation. The training group was used purely for constructing the visual dictionary — its subjects were never seen in the development and evaluation groups. Experiments on all datasets were carried out on face images which are closely cropped and downsampled to a size of 64×64. Each image set contains three images. The number of matched pairs and mismatched pairs is the same (balanced), in order to prevent a bias towards one of the pair types.

For the LFW dataset, 620 pairs of image sets were generated, with 310 pairs for development group and 310 pairs for evaluation group. The generic subset from LFW view 1 was used for the training group.

For the CMU PIE dataset, we used the near frontal poses (C05, C07, C09, C27 and C29), resulting in 170 images per subject with various illuminations and expressions. We randomly selected 8 subjects for the training group while development and evaluation groups each have 30 subjects. 1,200 pairs of images were generated, with the development and evaluation groups having 600 pairs each.

The MOBIO dataset contains images captured from mobile devices. The quality of the images is generally poor with blurring from motion and smudged lenses, as well as changes in illumination between scenes. A Haar-based cascade classifier [32] was used to locate faces in each frame. The eyes within each face are located using a similar cascade classifier. If no eyes are located, their approximate location is inferred from the size of the face bounding box. The faces are then resized and cropped such that the eyes are centered with a 32-pixel inter-eye distance. We used the development subset of MOBIO, which contains 1,500 probe videos from 20 females and 27 males. We generated 832 pairs of images for the development group and 800 pairs for the evaluation group. The background data subset was used as the training group.

4.2. Comparative Performance Evaluation

The proposed approach is compared against several benchmark methods as well as recent state-of-the-art methods. The evaluated methods are representative techniques for exemplar-based and subspace-based approaches.

The exemplar-based techniques are: Laplacianface [15], Local Binary Pattern (LBP) [2], Multi-Region Histograms (MRH) [28], and Local Facial Features (LFF) [6]. The subspace-based techniques are: Mutual Subspace Method (MSM) [36], Kernel Affine Hull Method (KAHM) [5], and Local Principal Angle (Local-PA) [21].

We note that the above approaches can also be classified as either local or holistic in terms of the underlying feature extraction. LBP, MRH, LFF, Local-PA and the proposed approach are in the local based category, while Laplacianface, MSM and KAHM are in the holistic based category.

Similarity judgements in exemplar-based methods were carried out using the Modified Hausdorff Distance (MHD) [7]. The KAHM approach used a linear kernel with the parameters tuned according to the recommendations made in [5]. The best results are reported. For LBP, uniform histograms with $(8, 1)$ neighbourhoods are employed. The LBP block size was selected empirically as 7×9. In Laplacianface, the subspace dimensions were set by retaining enough leading eigenvectors to account for 98% of the overall energy in the eigen-decomposition. In Local-PA, the block size was 16×16, also obtained empirically.

Based on preliminary experiments, the proposed approach used the following parameters: the size of each region is 24×24, dimension of each DCT-based texture descriptor is 15, and the number of visual words in the dictionary is 1024.

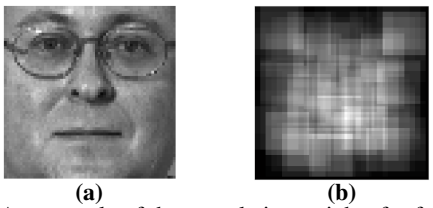
To generate compound descriptors, the distance between centers of simple cells (regions in the image plane) was selected as 4, 8 and 12. For images of size 64×64, this results in 1681 direct regions and 8153 compound regions. As four distance metrics are used for each local mode, the dimensionality of the resulting similarity vector for each image set pair is 39336. The discrimination performance appears to stabilise with a subset of 150 similarity features, as selected by the AdaBoost algorithm.

An example of cumulative weights of the most discriminant local modes obtained by the boosting algorithm is shown in Fig. 4. Cumulative weight for a pixel $I(x, y)$ is defined as the sum of the weights of the selected regions that include the pixel. Most of regions are selected from the inner part of the face, with stress on the regions around the mouth, nose and eyes.

The comparative results are shown in Table 1, with the verification accuracy defined as the average of the accuracy on matched and mismatched pairs. The relatively poor performance of the Laplacianface approach implies the difficulty of the recognition task, considering that the method is expected to perform relatively well if the imaging conditions do not differ greatly between training and test datasets.

The results show that in all experiments local approaches prevail over holistic techniques. This confirms the premise of this work: *relaxed local representations are more robust than rigid holistic representations*. Among exemplar-based methods, MRH and LFF outperform Laplacianface and LBP. Among the subspace approaches, Local PA outperforms MSM. We note that KAHM is marginally superior to MSM (with the exception of CMU-PIE), however LBP+KAHM significantly outperforms MSM for all experiments. This is consistent with the results reported in [5].

The proposed approach surpasses all other methods by a considerable margin on the LFW and PIE datasets. On LFW, the performance difference to LFF, the nearest competing approach, is 7.7 percentage points. On PIE, the improvement over the nearest method is close to 13 percentage points.



(a) **(b)**

Figure 4. An example of the cumulative weights for face regions selected by the boosting algorithm: **(a)** cropped face from PIE; **(b)** brighter regions correspond to higher cumulative weights.

Table 1. Average verification accuracy on LFW, PIE and MOBIO datasets. The methods are grouped into two categories: **(a)** exemplar based, and **(b)** subspace based. The proposed method uses both exemplar and subspace based similarity metrics.

	Method	**LFW**	**PIE**	**MOBIO**	overall
(a)	Laplacian [15] + MHD	65.48	69.17	85.50	73.38
	LBP [2] + MHD	79.35	78.17	94.75	84.09
	MRH [28] + MHD	86.45	75.50	96.75	86.23
	LFF [6]	88.06	78.17	97.75	87.99
(b)	MSM [36]	65.48	71.33	90.13	75.65
	Local-PA [21]	67.10	77.17	92.50	78.92
	KAHM [5]	66.13	67.83	90.38	74.78
	LBP + KAHM [5]	73.22	76.00	95.38	81.53
	Proposed method	**95.80**	**91.00**	**100.00**	**95.60**

5. Main Findings and Future Directions

We have proposed a novel image set matching technique for face verification, comprised of three aspects: **(i)** robust descriptors of face regions based on local features, partly inspired by the hierarchy in the human visual system, **(ii)** use of several subspace and exemplar metrics to compare corresponding face regions, **(iii)** jointly learning which regions are the most discriminative while finding the optimal mixing weights for combining metrics. Experiments on LFW, PIE and MOBIO face datasets show that the proposed algorithm obtains considerably better performance than several recent state-of-the-art techniques, such as Local Principal Angle and the Kernel Affine Hull Method.

We note that the region descriptors used in Section 2 somewhat resemble Sparse Representation (SR) and dictionary learning, as they are obtained through an overcomplete visual dictionary [8]. While SR methods usually utilise greedy algorithms like Matching Pursuit or convex optimisation [8] (which are computationally expensive), the descriptors here are obtained through closed-form equations. This is useful in large-scale data processing applications.

While the learning method presented here is specific to a verification system (ie. binary classification), extension to arbitrary M-class discrimination problems is possible. An M-class problem can be converted into a binary problem via the use of intra- and inter-personal spaces [23]. More specifically, instead of characterising class clusters, it is possible to characterise what kind of image variation is typical for the same person and what is for different persons. Theoretically this is achieved by training a binary classifier on the differences between two samples, ie. $\Delta = S_1 - S_2$. Based on learning the differences, two samples (here two sets obtained from a specific local region) are considered as representing the same person if they are classified as intra-personal variation. Conversely, two samples represent two unique individuals if their difference is classified as extra-personal variation.

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.

References

[1] Y. Adini, Y. Moses, and S. Ullman. Face recognition: the problem of compensating for changes in illumination direction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):721 –732, 1997.

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[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] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *IEEE Conf. Computer Vision and Pattern Recogition (CVPR)*, pages 2567–2573, 2010.

[6] S. Chen, S. Mau, M. Harandi, C. Sanderson, A. Bigdeli, and B. C. Lovell. Face recgonition from still images to video sequences: A local facial feature based framework. *EURASIP Journal on Image and Video Processing*, 2011.

[7] M. Dubuisson and A. Jain. A modified Hausdorff distance for object matching. In *Int. Conf. Pattern Recognition*, pages 566–568, 2004.

[8] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[9] A. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *IEEE Conf. Computer Vision and Pattern Recog.*, pages 26–36, 2003.

[10] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

[11] A. Hadid and M. Pietikäinen. Manifold learning for video-to-video face recognition. In *Lecture Notes in Computer Science*, volume 5707, pages 9–16, 2009.

[12] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Int. Conf. Machine learning*, pages 376–383, 2008.

[13] M. Harandi, C. Sanderson, S. Shirazi, and B. Lovell. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 2705–2712, 2011.

[14] M. T. Harandi, M. N. Ahmadabadi, and B. N. Araabi. Optimal local basis: A reinforcement learning approach for face recognition. *International Journal of Computer Vision*, 81(2):191–204, 2009.

[15] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using Laplacianfaces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[16] B. Heisele, P. Ho, J. Wu, and T. Poggio. Face recognition: component-based versus global approaches. *Computer Vision and Image Understanding*, 91(1-2):6–21, 2003.

[17] 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. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[18] T. Kanade and A. Yamada. Multi-subregion based probabilistic approach toward pose-invariant face recognition. In *IEEE Int. Symposium on Computational Intelligence in Robotics and Automation*, pages 954–959, 2003.

[19] T.-K. Kim, O. Arandjelović, and R. Cipolla. Boosted manifold principal angles for image set-based recognition. *Pattern Recognition*, 40(9):2475–2484, 2007.

[20] T. Lee. Image representation using 2D Gabor wavelets. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(10):959–971, 1996.

[21] X. Li, K. Fukui, and N. Zheng. Image-set based face recognition using boosted global and local principal angles. In *ACCV, Lecture Notes in Computer Science*, volume 5994, pages 323–332, 2010.

[22] S. Marcel, C. McCool, P. Matejka, T. Ahonen, J. Cernocky, et al. On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation. In *Lecture Notes in Computer Science*, volume 6388, pages 210–225, 2010.

[23] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000.

[24] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 84–91, 1994.

[25] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

[26] C. Sanderson. Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments. Technical report, NICTA, 2010.

[27] C. Sanderson, S. Bengio, and Y. Gao. On transforming statistical models for non-frontal face verification. *Pattern Recognition*, 39(2):288–302, 2006.

[28] C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Lecture Notes in Computer Science*, volume 5558, pages 199–208, 2009.

[29] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[30] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *IEEE Conf. Automatic Face and Gesture Recognition*, page 53, 2002.

[31] K. Tieu and P. Viola. Boosting image retrieval. *Int. J. Computer Vision*, 56(1-2):17–36, 2004.

[32] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Computer Vision*, 57(2):137–154, 2004.

[33] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[34] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[35] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *The Journal of Machine Learning Research*, 4:913–931, 2003.

[36] O. Yamaguchi, K. Fukui, and K.-i. Maeda. Face recognition using temporal image sequence. In *IEEE Conf. Automatic Face and Gesture Recognition*, pages 318–323, 1998.