ENSEMBLE OF FURTHEST SUBSPACE PAIRS FOR ENHANCED IMAGE SET MATCHING

Mehrtash T. Harandi, Conrad Sanderson, Abbas Bigdeli, Brian C. Lovell

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

ABSTRACT

Recently it has been shown that the performance of image set matching methods can be improved by clustering set samples into smaller and more coherent groups. Typically, set samples are treated independently during clustering, ie., clustering criteria have not been defined to exploit set characteristics. In this paper we introduce a novel approach to image set clustering by considering the similarities between subspaces instead of similarities between samples. We exploit an ensemble learning technique to create an ensemble of subspace pairs. Each pair has the property that its members are located at the furthest distance in the sense of distances between subspaces. Object recognition experiments on the CMU-MoBO and ETH-80 datasets show that the proposed method obtains higher discrimination accuracy in comparison to several benchmark methods as well as the recently proposed Kernel Affine Hull Method.

Index Terms— image set matching, object recognition, linear subspaces, principal angle, ensemble learning.

1. INTRODUCTION

In recent years, several studies in the field of computer vision have utilised set information for visual object recognition and categorisation, to obtain considerable improvements over the conventional single-image-to-single-image matching scenario [1, 5, 12, 14]. Exploiting set information is mainly driven by the need for superior discrimination accuracy in the presence of practical issues such as misalignment as well as variations in pose and illumination. Broadly speaking, studies in the field of multiple image matching can be categorised into three streams: **(i)** temporal model based, **(ii)** exemplar-based approaches, **(iii)** image set matching.

Temporal model based approaches exploit the consecutive nature of video frames. Dependency on temporal information is restrictive, as in many image set matching applications it may not be practical or even possible to extract objects of interest from consecutive frames. As temporal model based approaches are of little relevance to this paper, for brevity we will not address them further. Exemplar-based methods are typically direct extensions of single-to-single matching approaches and handle each image in a set independently.

The third category, ie., image set matching, benefits from linear structures and considers image sets as linear subspaces [1, 6, 12, 14]. Subspaces appear to be appropriate models for this task since they are able to accommodate the effects of various image variations. Recent findings suggest that creating smaller sets by clustering may improve recognition accuracy [4, 12, 13]. This can be partially explained by considering the fact that images usually lie on highly nonlinear and complex manifolds, and hence linear models like subspaces might not be able to capture the necessary structure for adequate discrimination.

The basic idea of typical clustering approaches is to define a similarity measure between single images, independently, and use it to arrange the images into clusters [4, 12]. Hadid et al. [4] used Euclidean distances between single images to approximate a linear model of images from their neighbouring points and then embedded the results on a lower dimensional manifold. Wang et al. [12] devised a clustering algorithm based on differences between geodesic and Euclidean distances, where again, both geodesic and Euclidean distances are obtained by treating images independently.

This paper aims to tackle the set clustering problem on a different ground. Towards this we propose to address the set clustering problem based on an ensemble learning technique. The proposed method utilises Random Subspace Method (RSM) [8] and the concept of principal angles to create an ensemble of subspace pairs. In each pair, the subspaces are located at the furthest distance ie., most dissimilar. Intuitively this is similar to representing classes with their boundaries instead of their center of gravity.

Unlike previous approaches, our proposal hence considers the set similarity measure as a criterion for clustering. Experiments on two recognition tasks, namely face and object recognition, show that not only the proposed method improves the performance of conventional techniques based on principal angles [14], but it can also compete and outperform the state-of-the-art Kernel Affine Hull Method [1].

We continue the paper as follows. The proposed technique is introduced in Section 2. In section 3, we compare the performance of our proposed approach with other methods on two public datasets (CMU-MoBo and ETH-80). Section 4 concludes the paper and suggests future directions.

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.

2. PROPOSED APPROACH

Given an image set $\boldsymbol{\zeta} \in \mathbb{R}^{n \times l}$, one can extract at maximum $\frac{l!}{k!(l-k)!}$ subspaces of rank k by selecting k images out of available l images (where $k < l$). Grassmannian analysis is a convenient way to analyse image sets modelled by subspaces [5]. Any subspace of rank k can be modelled by a point on a Grassmannian manifold, $G_{n,k}$. As a result, an image set of length l can be represented by a cloud of points or a distribution on $G_{n,k}$.

Unfortunately computing all available subspaces of a particular rank for an image set is intractable. Instead of describing an image set by all its subspaces, we propose to use a random ensemble of subspaces along their furthest points on $G_{n,k}$ (see Fig. 1 for a graphical interpretation). As a result, our method approximates the subspace cloud on $G_{n,k}$ by its boundary.

We continue this section by first describing the concept of pairs of furthest subspaces, followed by elaborating on ensemble of furthest pairs for image set matching.

2.1. Furthest Subspaces

Consider two bases $\Gamma = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_l\} \in \mathbb{R}^{n \times l}$ and $\Omega = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_k\} \in \mathbb{R}^{n \times k}$ are given, where n is the number of available pixels of an image. Without loss of generality, we assume $l > k$. We are interested in finding a subspace $\Psi = \{\boldsymbol{X}_s, \boldsymbol{X}_{s+1}, \cdots, \boldsymbol{X}_{s+k-1}\} \in \mathbb{R}^{n \times k}$ in Γ, ie., $\Psi \subset \Gamma$, where the distance between Ω and Ψ is maximised.

Similar to the definition of principal angles, we consider the angles between $\boldsymbol{X}_i$ and $\boldsymbol{Y}_j$ as the measure of similarity between two vectors in Γ and Ω. We define Ψ as the first k unique vectors of Γ that have the minimum of maximum similarity to subspace Ω, ie.,

$$s = \operatorname*{argmin}_i \left(\max_j \left(\langle \boldsymbol{X}_i, \boldsymbol{Y}_j \rangle \right) \right) \qquad (1)$$

The uniqueness implies that Ψ must be full-rank, ie., if $\boldsymbol{X}_s$ satisfies (1), it cannot be selected afterwards. We note that $\max_j \left(\langle \boldsymbol{X}_i, \boldsymbol{Y}_j \rangle \right)$ captures the maximum similarity between vector $\boldsymbol{X}_i$ and subspace Ω. Hence by selecting k vectors with minimum similarity, one can construct the most dissimilar or furthest subspace Ψ in the sense of principal angle.

We emphasise that if a pair of furthest subspaces is extracted from a particular image set, ie., images that generate subspaces Γ and Ψ belong to the same image set, then the pair of furthest subspaces can be considered as a clustering approach, where subspace similarities are taken into account as the clustering criterion. In other words, the subspaces Γ and Ψ cluster the image set into two clusters at the extreme.

2.2. Ensemble of Pairs of Furthest Subspaces

As mentioned before, while the available subspaces of particular rank may provide valuable information about the image set, considering all of them for classification is intractable. Since discovering useful subspaces for classification is N-P hard and very challenging, we propose to address the subspace derivation based on ensemble learning and the idea of furthest subspace pairs.

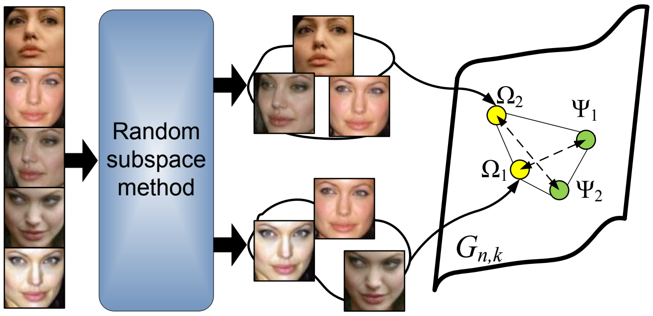

Fig. 1: Graphical interpretation of the proposed method. For any image-set, a random ensemble of subspaces (Ω_i) and their furthest subspaces (Ψ_i) are generated and used for classification.

In recent years various ensemble methods for object and face recognition have been proposed that show promising results [10, 15]. Among the various ensemble construction approaches, the Random Subspace Method (RSM) has received considerable attention in face recognition [15]. In RSM an ensemble of classifiers is constructed on independently and randomly selected subsets.

In this work, an image set is represented by an ensemble of furthest subspace pairs. This is achieved by random selection of k images out of an image set and using the remaining images to construct the furthest subspace. To compare two image sets, we compare their subspaces based on the concept of principal angles [14], followed by aggregating the results. If $\boldsymbol{\zeta}_1 \in \mathbb{R}^{n \times l_1}$ and $\boldsymbol{\zeta}_2 \in \mathbb{R}^{n \times l_2}$ are two linear subspaces in $\mathbb{R}^n$ with minimum rank $r = \min(\operatorname{rank}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2))$, then there are exactly r uniquely defined principal angles between $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$:

$$\cos(\theta_i) = \max_{\boldsymbol{x}_i \in \boldsymbol{\zeta}_1, \, \boldsymbol{y}_j \in \boldsymbol{\zeta}_2} \boldsymbol{x}_i^T \boldsymbol{y}_j \qquad (2)$$

subject to $\boldsymbol{x}_i^T \boldsymbol{x}_i = \boldsymbol{y}_i^T \boldsymbol{y}_i = 1, \boldsymbol{x}_i^T \boldsymbol{x}_j = \boldsymbol{y}_i^T \boldsymbol{y}_j = 0, i \neq j$.

In Eqn. (2), $\theta_i \in [0, \pi/2]$ is the principal angle between the two subspaces and $i \in \{1, 2, \cdots, r\}$. To compare two subspaces, the cosine of the first principal angle or the Max-Correlation metric [5] is used here.

There are several methods to aggregate the output of ensemble members. In order to keep the system as simple as possible, we used the majority voting scheme in our experiments. If a binary vector T_i for subspace i is defined as:

$$T_i(\boldsymbol{X}, j) = \begin{cases} 1 & \text{if label of closest subspace to } \boldsymbol{X} \text{ is } j \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

then the majority voting rule can be expressed by:

$$MV(\boldsymbol{X}) = \operatorname*{argmax}_j \left(\sum\nolimits_i^{N_{\text{subspace}}} T_i(x, j) \right) \qquad (4)$$

Putting all together, the proposed algorithm can be described by the pseudo code shown in Fig. 2.

```
input:
  • training data $S = \{(\boldsymbol{\zeta}_i, y_i)\}_{i=1}^{N_{\texttt{training}}}$, where $\boldsymbol{\zeta}_i = \left\{\boldsymbol{I}_1^i, \boldsymbol{I}_2^i, \cdots, \boldsymbol{I}_{l_i}^i\right\} \in \mathbb{R}^{n \times l_i}$ is the image set
    and $y_i \in \{1, 2, \cdots, c\}$ are the corresponding class labels
  • number of random subspaces per set, ie., $N_{\texttt{subspace}}$ and number of images per random subspace, ie., $k$
  • probe set $\boldsymbol{P}$ whose label should be determined
$\Lambda = \emptyset$
for all sets $\{(\boldsymbol{\zeta}_i, y_i)\}_{i=1}^{N_{\texttt{training}}}$ do
  for $j = 1$ to $N_{\texttt{subspace}}$ do
    Randomly select $k$ images from $\boldsymbol{\zeta}_i = \left\{\boldsymbol{I}_1^i, \boldsymbol{I}_2^i, \cdots, \boldsymbol{I}_{l_i}^i\right\}$.
    Create a basis by applying QR decomposition to the randomly selected images and call the Q part $\boldsymbol{\Omega}_i^j$.
    Generate the furthest subspace $\boldsymbol{\Psi}_i^j$ to $\boldsymbol{\Omega}_i^j$ using the remaining images of $\boldsymbol{\zeta}_i$.
    Add $\left(\boldsymbol{\Omega}_i^j, y_i\right)$ and $\left(\boldsymbol{\Psi}_i^j, y_i\right)$ to the ensemble $\Lambda$, ie., $\Lambda = \Lambda \bigcup \left\{\left(\boldsymbol{\Omega}_i^j, y_i\right)\right\} \bigcup \left\{\left(\boldsymbol{\Psi}_i^j, y_i\right)\right\}$.
  end for
end for
for $j = 1$ to $N_{\texttt{subspace}}$ do
  Randomly select $k$ images from $\boldsymbol{P}$.
  Create a basis by applying QR decomposition to the randomly selected images and call the Q part $\boldsymbol{\Omega}_P^j$.
  Generate the furthest subspace $\boldsymbol{\Psi}_P^j$ to $\boldsymbol{\Omega}_P^j$ using the remaining images of $\boldsymbol{P}$.
end for
Compare all $\boldsymbol{\Psi}_P^j$ and $\boldsymbol{\Omega}_P^j$ to the members of ensemble $\Lambda$ using the maximum correlation similarity measure and use majority voting
aggregation method (Eqns. (3) and (4)) to derive the label of $\boldsymbol{P}$.
output:
  • the label of probe set $\boldsymbol{P}$
```

Fig. 2: Pseudo code for the proposed algorithm.

3. EXPERIMENTS

The proposed approach was compared and contrasted to previous state-of-the-art and benchmark methods on two recognition tasks, namely face and object recognition. We used the CMU-MoBo [3] and ETH-80 [9] datasets. For both datasets we randomly generated ten splits of training and testing sets. All the images in our experiments were cropped and normalised to a size of 32 × 32 pixels.

CMU-MoBo consists of motion sequences of 25 people walking on a treadmill. For each person video recordings were made for 4 walking styles (slow walk, fast walk, inclined walk and slow walk while holding a ball), viewed from a set of fixed cameras. Sample images of this dataset are shown in Fig. 3. For each split, we randomly considered one of the available styles as training data and the remaining styles were used for testing. Each set in a training or testing sequence was represented by sixty randomly selected images.

ETH-80 contains images of eight object categories: apples, cows, cups, dogs, horses, pears, tomatoes, and cars. Each category includes ten object subcategories (eg., various dogs) in 41 orientations, resulting in 410 images per category. Examples are shown in Fig. 4. In each split, we considered one of the subcategories as training data and used the remaining subcategories as test data. As a result each set contained 41 images.

The proposed approach was compared against Eigenface [11], Laplacianface [7], Mutual Subspace Method (MSM) [14], and the recently proposed Kernel Affine Hull Method (KAHM) [1]. These are representative techniques for exemplar-based and subspace-based approaches.

Similarity judgements in Eigenface [11] and Laplacianface [7] methods were carried out using the original Hausdorff distance (HD) as well as the modified version (MHD) [2]. Given two sets of points A and B as well as an underlying point to point distance $\|.\|$, the HD is defined as:

$$d_{\text{HD}}(A, B) = \max\left(\max_{a \in A}\min_{b \in B}\|a - b\|\,,\ \max_{b \in B}\min_{a \in A}\|a - b\|\right) \quad (5)$$

Intuitively, if the HD is q, then every point of A must be within a distance q of some point B, and vice versa. For image processing applications, MHD was shown to be more robust against outliers. The MHD is defined as:

$$d_{\text{MHD}}(A, B) = \max\left(\frac{\sum_{a \in A}\min_{b \in B}\|a - b\|}{|A|}, \frac{\sum_{b \in B}\min_{a \in A}\|a - b\|}{|B|}\right) \quad (6)$$

where $|A|$ is the cardinality of set A.

In our experiments the kernel function for KAHM was linear and KAHM's parameters were tuned according to the recommendations made in [1]; the best results are reported. In Eigenface and Laplacianface, subspace dimensions were set by retaining enough leading eigenvectors to account for 98% of the overall energy in the eigen-decomposition.

The recognition accuracies for all methods are shown in Table 1. The relatively poor performance of the Eigenface-HD and Laplacianface-HD subspaces implies the difficulty of

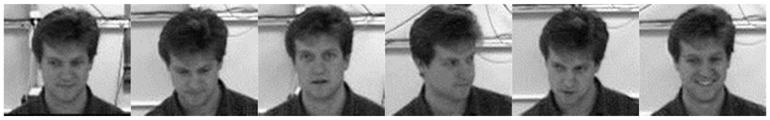

Fig. 3: Examples of variations in the CMU-MoBo dataset.

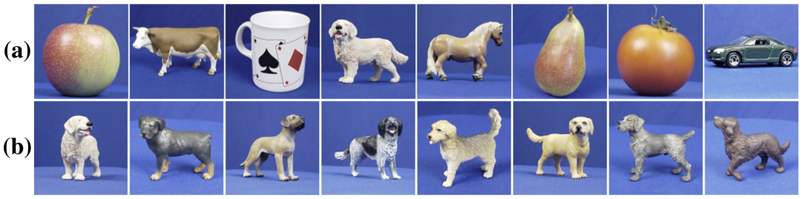

Fig. 4: **(a)** examples from the 8 object categories in the ETH-80 dataset; **(b)** examples of various classes within an object category.

Table 1: Average accuracy on the CMU-MoBo and ETH8-80 datasets. The standard deviation is shown in brackets.

Method	**CMU-MoBo**	**ETH-80**
Eigenface-HD	42.63 (4.8)	55.00 (5.4)
Eigenface-MHD	61.40 (2.6)	61.67 (4.9)
Laplacianface-HD	32.81 (4.9)	57.36 (9.2)
Laplacianface-MHD	55.79 (5.5)	59.44 (7.4)
MSM [14]	50.70 (3.8)	60.83 (5.0)
KAHM [1]	62.81 (8.5)	53.47 (4.9)
Proposed method	**64.38 (6.0)**	**73.19 (5.8)**

the recognition task, considering that both can be expected to perform relatively well if the imaging conditions do not greatly differ between training and test data sets.

The proposed approach attains better performance than all of the other methods. In comparison to the nearest competitor on the CMU-MoBo dataset (KAHM), the proposed method obtains accuracy that is about 1.6 percentage points higher. On the ETH-80 dataset, the difference in performance to the nearest competitor is considerably more prominent. Specifically, the proposed method obtains accuracy that is about 11.5 percentage points higher than obtained by Eigenface-MHD.

We also note that MSM outperforms KAHM on the ETH-80 dataset. This drop in KAHM's performance might be due to the presence of the background. In contrast to the face recognition task, where images were cropped from the internal part of the face, ETH-80 images contained the background.

4. MAIN FINDINGS & FUTURE DIRECTIONS

We have described a novel approach to image set matching by exploiting ensemble of subspaces. Towards this, we proposed how semi-random ensembles of subspaces can be generated and how the ensembles can be compared to address the image set matching problem. This was achieved by defining the concept of furthest subspace pair based on the principal angle.

The proposed method was compared with several benchmark methods, as well as a recently proposed state-of-the-art approach, on the CMU-MoBo and ETH-80 datasets. The empirical evaluations demonstrate promising improvements in recognition accuracy.

We intend to extend the proposed concept by pruning the generated ensemble. Pruning the overall ensemble not only aims to remove the redundancy but it also targets the poor subspaces that may deteriorate the performance of overall ensemble.

5. REFERENCES

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

[2] M.-P. Dubuisson and A. Jain. A modified Hausdorff distance for object matching. In *Int. Conf. Pattern Recognition*, pages 566–568, 1994.

[3] R. Gross and J. Shi. The CMU Motion of Body (MoBo) Database. Technical Report CMU-RI-TR-01-18, CMU Robotics Institute, 2001.

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

[5] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *International Conference on Machine Learning (ICML)*, pages 376–383, 2008.

[6] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. 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.

[7] 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.

[8] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Trans. PAMI*, 20(8):832 –844, 1998.

[9] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 409–415, 2003.

[10] J. Lu, K. Plataniotis, A. Venetsanopoulos, and S. Li. Ensemble-based discriminant learning with boosting for face recognition. *IEEE Transactions on Neural Networks*, 17(1):166–178, 2006.

[11] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

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

[13] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 74–81, 2011. DOI: 10.1109/CVPRW.2011.5981881.

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

[15] Y. Zhu, J. Liu, and S. Chen. Semi-random subspace method for face recognition. *Image and Vision Computing*, 27(9):1358–1370, 2009.