RESEARCH ARTICLE

Video based object representation and classification using multiple covariance matrices

Yurong Zhang1,2*, Quan Liu1

1 Wuhan University of Technology, School of Information Engineering, Wuhan, China, 2 HuiShang Vocational Technical College, Electronics Information Department, Hefei, China

* zhangyurong815@163.com

Abstract

Video based object recognition and classification has been widely studied in computer vision and image processing area. One main issue of this task is to develop an effective representation for video. This problem can generally be formulated as image set representation. In this paper, we present a new method called Multiple Covariance Discriminative Learning (MCDL) for image set representation and classification problem. The core idea of MCDL is to represent an image set using multiple covariance matrices with each covariance matrix representing one cluster of images. Firstly, we use the Nonnegative Matrix Factorization (NMF) method to do image clustering within each image set, and then adopt Covariance Discriminative Learning on each cluster (subset) of images. At last, we adopt KLDA and nearest neighborhood classification method for image set classification. Promising experimental results on several datasets show the effectiveness of our MCDL method.

1 Introduction

With the recent development in imaging techniques, multiple images of an object are usually available in many cases, such as video based surveillance, multi-view camera networks, etc. Object recognition from these multiple images is formulated as an image set (video) classification problem and has attracted more and more interests and attention in computer vision and machine learning area in recent years [1, 2, 3, 4, 5, 6, 7]. This technique can be widely used in many computer vision problems. For example, in visual object search task [8,9,10], one can use multiple images to retrieve and recognize the similar visual objects. In face recognition problem [11], we can also use multiple face images to conduct person identification. Compared with the traditional single image based object recognition and learning, video model generally contains more visual appearance contents and thus performing more robustly and effectively on image set representation [12, 13, 14, 15, 16, 17, 18].

One of the main problems and challenges for video based object recognition is to develop an effective method to represent an image set or sequence. In recent years, many methods have been proposed for image set representation and classification. Other main problems include image set classifier development, image set clustering methods and so on. In this paper, we focus on image set representation. Kim et al. [19] proposed Discriminant-analysis of...
Canonical Correlations (DCC) to represent an image set by using single linear subspace. Hamm et al. [3] proposed Grassmann Discriminant Analysis (GDA) which uses multiple local linear subspaces to represent an image set. Besides linear subspace, nonlinear subspace methods have also been used for image set representation. For nonlinear subspace based representation, Wang et al. [20] presented an image set with nonlinear manifolds and used Manifold-Manifold Distance (MMD) method for image set representation and classification. Wang et al. [21] also proposed Manifold Discriminant Analysis (MDA) to obtain a more discriminative feature space to represent a set of images. In additional to above methods, probabilistic models have also been used for image set representation and classification. Shakhnarovich et al. [4] used a single Gaussian model for set modeling. Arandjelovic et al. [1] further provided a method to use Gaussian Mixture Models (GMM) to image set representation. Wang et al. [11] proposed Discriminant Analysis on Riemannian Manifold of Gaussian Distributions (DARG) to learn a discriminative representation for image set.

As one of the probabilistic methods, Covariance Discriminative Learning (CDL) [5] has been widely used for image set representation. The core idea of CDL method is to represent an image set using a single covariance matrix. One benefit of CDL representation is that it makes no assumption about the set data distribution and thus providing a simple and effective representation for an image set with any kinds of features. However, when data samples are drawn from a union of multiple subspaces, traditional CDL generally fails to provide an accurate and reliable representation.

In this paper, we present a new image set representation method called multiple covariance discriminative learning (MCDL), which aims to represent an image set using multiple covariance matrices with each covariance matrix representing one cluster of images. Comparing with previous single CDL method [5], MCDL explores the data distribution of multiple subspaces more thus providing a more faithful representation. To do that we first use the Non-negative Matrix Factorization (NMF) technique to cluster the samples into their respective subspaces. Then, we adopt Covariance Discriminative Learning (CDL) on representing each cluster (subset) of images which lies in a single subspace, as shown in Fig 1. Note that covariance-based visual representation has been used in many applications [22, 23]. Different from these works, here we focus on multiple covariance matrices representation, which considers multi-subspaces property of image set data and thus providing a more effective descriptor for image set data. For set classification, we first define a method to measure the similarity between image sets based on MCDL and then adopt KLDA and nearest neighborhood classification method [5] for image set classification. Experimental results on several datasets show the effectiveness and benefits of the proposed MCDL method.

The remainder of this paper is organized as follows. In the materials and methods part, we introduce nonnegative matrix factorization (NMF) data clustering method and propose our Multiple Covariance Matrices representation and Kernel LDA classification method. At last, we apply MCDL method to some datasets to evaluate the effectiveness of the method.

2 Materials and methods

The experimental data in our study was acquired legitimately from international standard database, and this study was approved by the Local Ethics Committee of Wuhan University of Technology.

2.1 NMF clustering

Nonnegative Matrix Factorization (NMF) [24, 25] is a matrix factorization algorithm that has been widely used in many machine learning problems. Let $X = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{p \times n}$ be $n$ data
points in $p$-dimensional space. The aim of NMF is to find two smaller nonnegative matrices $F \in \mathbb{R}^{p \times k}$ and $G \in \mathbb{R}^{n \times k}$ whose product can approximate the original matrix $X$ as close as possible, i.e.,

$$X \simeq FG^T.$$  \hspace{1cm} (1)

Using Euclidean distance (or Frobenius norm) residual function, the above approximation problem can be formulated as the following optimization,

$$\min_{F,G} \|X - FG^T\|_F^2 = \sum_i (x_i - Fg_i)^2$$  \hspace{1cm} (2)

$$\text{s.t. } F_y \geq 0, \ G_y \geq 0.$$  \hspace{1cm} (3)

From optimization aspect, although the above objective functions are convex in $F$ or $G$ only, there are not convex in both of this two variables. Thus, it is difficult to develop an algorithm to find the global optimal solution for this problem.

Lee and Seung [18] has presented an effective update algorithm which iteratively updates the current solution as follows,

$$F_{ik} \leftarrow F_{ik} \frac{(XG)_{ik}}{(FG)^{T}_{ik}}, \ G_{ik} \leftarrow G_{ik} \frac{(X^TF)_{ik}}{(G^TF)_{ik}}$$  \hspace{1cm} (4)

It has been proven that the above update algorithm can converge to a local optimal solution.

**Fig 1.** The general flow of our MCDL method.

https://doi.org/10.1371/journal.pone.0176598.g001
The above NMF model has been widely used in many applications. One important aspect of NMF is that it can be used for data clustering. In fact, let \( F^* = (f_1^*, f_2^*, \cdots, f_k^*) \) and \( G^* \) be the optimal solution of the above optimization problem. Then, \( f_i^*, i = 1 \cdots k \) can be regarded as the cluster centroid, and the optimal \( G_{ik}^* \) can be viewed as the continuous coefficient of data \( x_i \) belonging to cluster \( c_k \). In clustering process, we can use the maximum coefficient of \( G_{ik}^* \) to determine the cluster label of data \( x_i \).

### 2.2 Image set modeling with multiple covariance matrices

In this section, an effective method is proposed to represent image sets by using multiple covariance matrices. Based on this representation, a similarity metric between two image sets is further computed.

**A. Image set representation.** We first propose an effective image set representation by using multiple covariance matrices, called Multiple Covariance Discriminative Learning (MCDL). Formally, given a video (or image set) \( X = (x_1, x_2, \cdots, x_n) \), we first use the above NMF method to do clustering on \( X \) and obtain clustering results \( X = \{X_1, X_2, \cdots, X_k\} \) with \( k \) clusters. Here, \( X_i \) is the image subset belonging to the \( i \)-th cluster.

Then, each cluster \( X_i \) is represented with a \( d \times d \) covariance matrix as follows,

\[
C_i = \frac{1}{n-1} \sum_{h=1}^{n} (X_i(h) - \bar{X}_i)(X_i(h) - \bar{X}_i)^T.
\]

(5)

where \( \bar{X}_i \) is the mean of the \( i \)-th cluster, and \( X_i(h) \) is the \( h \)-th element in cluster \( X_i \).

At last, the whole image set \( X \) can be represented by using a set of covariance matrices as follows,

\[
C = \{ C_1, C_2, \cdots, C_k \}.
\]

(6)

when \( k = 1 \), our MCDL degenerates to the traditional CDL method [5]. Therefore, the proposed MCDL can be regarded as a general extension of CDL representation. Comparing with CDL, MCDL can represent the variations of images in an image set more sufficiently and effectively while maintaining the benefit of CDL representation.

**B. Similarity metric for MCDL.** Based on the above MCDL representation, we propose a method to define a similarity metric between two image sets \( S_1 \) and \( S_2 \) whose covariance matrix representations are \( C_1 \) and \( C_2 \) respectively. Formally, let \( C_1 = \{ C_1^1, C_1^2, \cdots, C_1^k \} \) and \( C_2 = \{ C_2^1, C_2^2, \cdots, C_2^k \} \), it is known that for any covariance matrix \( C_i^h \) or \( C_i^k \), it is symmetric positive definite (SPD). For any SPD matrix, it does not lie in a Euclidean space but on the Riemannian manifold [5, 26]. Therefore, it is necessary to map \( C_i^h \) or \( C_i^k \) from Riemannian manifold to Euclidean space using the following logarithm operator [5],

\[
\Psi_{M \rightarrow E} : C \rightarrow \log (C)
\]

(7)

where

\[
\log (C) = \{ \log (C_1^1), \log (C_2^1), \cdots, \log (C_k^1) \}.
\]

Using this mapping \( \Psi_{M \rightarrow E} \), the similarity between two image sets \( S_1 \) and \( S_2 \) can be defined as the following three main steps.
Step 1. Compute the similarity between covariance matrices $C_i^1$ and $C_j^2$ as the inner product between them, i.e.,

$$K_{ij} = k(C_i^1, C_j^2) = \text{Tr}(\log(C_i^1) \cdot \log(C_j^2))$$

where $\text{Tr}(A)$ is the trace norm function of matrix $A$.

Step 2. Compute the optimal mapping $f$ between two covariance matrix set $C^1$ and $C^2$ by solving the following optimization

$$\min_f \sum_{i=1}^{k} \sum_{j=0}^{k} f_{ij} K_{ij}$$

$$s.t. \sum f_{ij} = 1, \sum f_{ij} = 1, f_{ij} \in \{0, 1\}$$

The above problem is known as bipartite graph matching problem and can be efficiently and effectively solved by using Hungarian algorithm.

Step 3. Calculate the mapping similarity between covariance matrix set $C^1$ and $C^2$ as follows,

$$K(C^1, C^2) = \sum_{i=1}^{c} \text{Tr}(\log(C_i^1) \cdot \log(C_f^j))$$

Note that the above similarity function $K$ is the combination of linear kernel functions $k(C_i^1, C_j^2)$. Therefore it is also a desired kernel function.

### 2.3 Image set classification

Based on the above MCDL representation and associated metric definition method, we can provide an effective classification method for image set. Generally, our classification method contains two main steps. Firstly, we use the Kernel Linear Discriminant Analysis (KLDA) [27, 5] method to extract a kind of discriminative feature for our MCDL representation. Then, we use nearest neighbor classification method to do classification on image set data. Let $X = \{X_1, X_2, \cdots, X_m\}$ be $m$ image sets belonging to $c$ classes. For each pair of set $X_i$ and $X_j$, we extract the MCDL representations for them and then compute the similarity kernel function $K(i,j)$ between them. The aim of KLDA is to solve the following optimization,

$$\max_p \frac{\text{tr}(p^T KLKp)}{p^T K K p}$$

$$s.t. p^T p = 1$$

where $K$ is the kernel matrix which is computed using Eq (10), and $L$ is the class label matrix defined as,

$$L_{ij} = \begin{cases} 
1 & \text{if } X_i, X_j \text{ belong to the } k \text{-th class} \\
\frac{1}{m_k} & \text{otherwise.} 
\end{cases}$$

where $m_k$ is the number of data points belonging to class $k$ and $\sum_{k=1}^{c} m_k = m$. It is well known that optimal solution $p$ can be obtained by computing the eigenvector corresponding to the largest eigenvalue. By further grouping the first largest $(c-1)$ eigenvectors, we can obtain $P = [p_1, \cdots, p_{c-1}]$ and get the $c-1$ projected feature vector by $z = P^T K$, where $K = [K_{11}, \cdots, K_{mm}]$. After KLDA projection, we then use nearest neighbor classification to classify image sets [5].
3 Experiments and results

In this section, we implement and test our MCDL method on several datasets to evaluate the effectiveness and benefits of our method. The detail introduction of these datasets are given below. These datasets have been widely used in many other methods. We have compared our MCDL method with some other methods including traditional Covariance Discriminative Learning (CDL) [5], Set to Set Distance Metric Learning (SSDML) [28], Manifold Discriminant Analysis (MDA) [21], Manifold-Manifold Distance (MMD) [20] and Discriminant Canonical Correlations [19].

3.1 Datasets description

- CMU MoBo [29] dataset contains 96 sequences of 24 persons. Each video contains approximately 300 frames.
- YTC [30] dataset contains 1910 face videos of 47 subjects. Each video contains several hundreds of frames.
- Cambridge-Gesture [31] dataset has 900 video sequences of 9 gestures in whole. Each gesture contains 100 videos. We divide it into five sets.
- ETH-80 [32] dataset contains 8 object categories in whole. Each category contains 10 object subcategories.

All the images used in these four datasets have been resized to the same $20 \times 20$ intensity images to make the same consistent dimension.

3.2 Classification results

We conduct the classification experiments for three cases by randomly selecting 50%, 70%, 90% image sets respectively for gallery and the rest image sets for probe. For fair comparison, the important parameters of each method were empirically tuned according to the recommendations in the original references. Fig 2 shows the classification results of all methods on the four datasets.

Here, we can observe that (1) CDL can return better performance in general, which indicates the effectiveness and benefits of the CDL method on conducting image set classification tasks. (2) MCDL obtains obvious better performance than other methods on gesture. Because in this dataset, the image in each set are usually lied on multiple subspaces. As discussed before, MCDL method performs more effectively and suitably for the data lying on multiple subspaces because it uses multiple covariance matrices representation instead of traditional single model representations such as MMD, MDA and DCC. (3) MCDL outperforms traditional CDL method and obtains the best performance on the four datasets. This clearly indicates the robustness and effectiveness of the proposed MCDL method on conducting image set representation and classification task. (4) MCDL obviously outperforms CDL on Gesture dataset. For this dataset, the variations of images in each set are very large due to different gestures and thus can be divided into several clusters. In this case, the proposed MCDL can capture these variations more effectively and sufficiently than single CDL.

We also test our method on the standard setting, which is summarized as follows. For each person in CMU MoBo [29] dataset, one set is used for gallery and the rest for probe. In YTC [30] dataset, we randomly chose 3 sets for gallery and 6 sets for probe. For Cambridge-Gesture [31] dataset, the first set for gallery and the rest four sets are used for probe. For each category in ETH-80 [32] dataset, five objects are selected for gallery and the rest 5 objects for probe. The
results are summarized in Table 1. It can be seen that our method can return better performance than other compared methods, which further demonstrates the robustness of the proposed MCDL method on image set classification tasks.

The classification accuracy of different methods on four datasets are summarized in Table 1. Here, we can note that comparing with other methods, CDL can return better performance, which indicates the effectiveness and benefits of the CDL method. Our MCDL generally outperforms traditional CDL method and obtains the best performance. This clearly indicates the robustness and effectiveness of the proposed MCDL method on conducting image set representation and classification task.

4 Conclusion

In this paper, we present a new image set representation method called multiple covariance discriminative learning (MCDL). The aim of MCDL is to represent an image set using multiple covariance matrices and each covariance matrix represents one cluster of images. To do that, firstly we use a Nonnegative Matrix Factorization (NMF) to conduct image clustering within each image set. Then, we adopt Covariance Discriminative Learning (CDL) to represent each

| Methods | CMUMoBo | YTC | Gesture | ETH-80 |
|---------|---------|-----|---------|--------|
| MMD     | 0.90    | 0.63| 0.10    | 0.86   |
| MDA     | 0.94    | 0.65| 0.11    | 0.89   |
| CDL     | 0.94    | 0.70| 0.69    | 0.97   |
| SSDML   | 0.24    | 0.82| 0.17    | 0.75   |
| DCC     | 0.88    | 0.65| 0.15    | 0.91   |
| MCDL    | 0.92    | 0.83| 0.94    | 0.98   |

https://doi.org/10.1371/journal.pone.0176598.t001
cluster (subset) of images. In terms of classification, we first define a method to measure the similarity between image sets based on MCDL and then adopt KLDA and nearest neighborhood classification method for image set classification. Experimental results show the effectiveness and benefits of the proposed method.

Supporting information
S1 Fig. The general flow of our MCDL method. (TIF)
S2 Fig. Classification accuracies of different methods on different datasets. (TIF)
S1 Table. Average accuracies of different methods on four datasets. (DOCX)

Author Contributions
Conceptualization: YZ QL.
Formal analysis: YZ.
Funding acquisition: YZ.
Investigation: YZ.
Methodology: YZ QL.
Supervision: QL.
Validation: YZ.
Writing – original draft: YZ.
Writing – review & editing: YZ QL.

References
1. O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, Face recognition with image sets using manifold density divergence, in: Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Computer Society Conference on, Vol. 1, IEEE, pp. 581–588 (2005).
2. L. Chen, Dual linear regression based classification for face cluster recognition, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE, pp. 2673–2680 (2014).
3. J. Hamm, and D. D. Lee, Grassmann discriminant analysis: a unifying view on subspace-based learning, in: Proceedings of the 25th international conference on Machine learning, ACM, pp. 376–383 (2008).
4. Shakhnarovich G., Fisher J. W., and Darrell T., Face recognition from long-term observations, in: Computer VisionECCV 2002, Springer, pp. 851–865 (2002).
5. R. Wang, H. Guo, L. S. Davis, and Q. Dai, Covariance discriminative learning: A natural and efficient approach to image set classification, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, pp.2496–2503 (2012).
6. O. Yamaguchi, K. Fukui, and K.-i. Maeda, Face recognition using temporal image sequence, in: Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, IEEE, pp. 316–323 (1998).
7. M. Yang, P. Zhu, L. Van Gool, and L. Zhang, Face recognition based on regularized nearest points between image sets, in: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, pp. 1–7 (2013).
8. Guan T., Wang Y., Duan L., and Ji R., On-device mobile landmark recognition using binarized descriptor with multifeature fusion, ACM Transactions on Intelligent Systems and Technology 7 (1) 1–29 (2015).

9. Y. Zhang, T. Guan, L. Duan, B. Wei, and J. Mao, Inertial sensors supported visual descriptors encoding and geometric verification for mobile visual location recognition applications, Signal Processing, pp. 17-26 (2015).

10. Ji R., Duan L.Y., Chen J., Huang T., and Gao W., Mining compact bag-of-patterns for low bit rate mobile visual search, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 23 (7) 3099–113 (2014). https://doi.org/10.1109/TIP.2014.2324291 PMID: 24835227

11. W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen, Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2048–2057 (2015).

12. H. Cevikalp, and B. Triggs, Face recognition based on image sets, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, pp. 2567–2573 (2010).

13. S. Chen, A. Wiliem, C. Sanderson, and B. C. Lovell, Matching image sets via adaptive multi convex hull, in: Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, IEEE, pp. 1074–1081 (2014).

14. Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen, Image sets alignment for video-based face recognition, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, pp. 2626–2633 (2012).

15. Y. Hu, A. S. Mian, and R. Owens, Sparse approximated nearest points for image set classification, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, pp. 121–128 (2011).

16. K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, Vol. 1, IEEE, pp. I–313 (2003).

17. Nishiyama M., Yamaguchi O., and Fukui K., Face recognition with the multiple constrained mutual subspace method, in: Audio-and Video-Based Biometric Person Authentication, Springer, pp. 71–80 (2005).

18. Ji R., Cao L., and Wang Y., Joint depth and semantic inference from a single image via elastic conditional random field, Pattern Recognition, pp. 2658–281 (2016).

19. Kim T.-K., Kittler J., and Cipolla R., Discriminative learning and recognition of image set classes using canonical correlations, Pattern Analysis and Machine Intelligence, IEEE Transactions on 29 (6) 1005–1018 (2007).

20. R. Wang, S. Shan, X. Chen, and W. Gao, Manifold-manifold distance with application to face recognition based on image set, in: Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on, IEEE, pp. 1–8 (2008).

21. R. Wang, X. Chen, Manifold discriminant analysis, in: Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on, IEEE, pp. 429–436 (2009).

22. J. Zhang, L. Wang, and L. Zhou, Exploiting structure sparsity for covariance-based visual representation, arXiv: 1610.08619.

23. J. Ren, and X. Wu, Bidirectional covariance matrices: A compact and efficient data descriptor for image set classification, Intelligence Science and Big Data Engineering. Image and Video Data Engineering.

24. D. D. Lee, and H. S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems (2001)

25. C. Ding, X. He, and H. D. Simon, On the equivalence of nonnegative matrix factorization and spectral clustering, in: Proc. SIAM Data Mining Conf (2005).

26. Arsigny V., Fillard P., Pennec X., and Ayache N., Geometric means in a novel vector space structure on symmetric positive-definite matrices, SIAM journal on matrix analysis and applications 29 (1) 328–347 (2007).

27. Baudat G., and Anouar F., Generalized discriminant analysis using a kernel approach, Neural computation 12 (10) 2385–2404 (2000). PMID: 11032039

28. P. Zhu, L. Zhang, W. Zu, and D. Zhang, From point to set: Extend the learning of distance metrics, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, pp. 2664–2671 (2013).

29. R. Gross, and J. Shi, The cmu motion of body (mobo) database.

30. M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, Face tracking and recognition with visual constraints in real-world videos, in: Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on, IEEE, pp. 1–8 (2008).
31. T.-K. Kim, K.-Y. K. Wong, and R. Cipolla, Tensor canonical correlation analysis for action classification, in: Computer Vision and Pattern Recognition (CVPR), 2007 IEEE Conference on, IEEE, pp. 1–8 (2007).

32. B. Leibe, and B. Schiele, Analyzing appearance and contour based methods for object categorization, in: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, Vol. 2, IEEE, pp. II–409 (2003).