# MULTI-VIEW DEEP METRIC LEARNING FOR IMAGE CLASSIFICATION

*Dewei Li*[1,2] *Jingjing Tang*[1,2] *Yingjie Tian*[2,3,*] *Xuchan Ju*[4,5]

[1]University of Chinese Academy of Sciences, Beijing 100049, China
[2]Research Center on Fictitious Economy & Data Science,
Chinese Academy of Sciences, Beijing 100190, China
[3] Key Laboratory of Big Data Mining and Knowledge Management,
Chinese Academy of Sciences, Beijing 100190, China
[4]School of Economics and Management, Tsinghua University, Beijing 100083, China
[5]Postdoctoral Programme of Agricultural Bank of China, Beijing 100005, China.

## ABSTRACT

In this paper, we propose a new deep metric learning approach, called MVDML, for multi-view image classification. Multi-view features can provide more information than single view, however, it is a challenge to exploit and fuse the complementary information from multiple views. Multiple deep neural networks are constructed, each corresponds to a view, to extract nonlinear information from images. The nonlinear transformation is an improvement on linear transformation of metric learning. All the original images will be transformed into a lower-dimensional space. In each new space, the difference between intra-class distance and inter-class distance is maximized. To extract information from different views as much as possible, the difference between different views of the same image is minimized. The numerical experiments verify that our model can obtain competitive performance in image classification and runs faster than the baseline methods.

***Index Terms***— Metric learning; Multi-view learning; Deep learning; Neural network

## 1. INTRODUCTION

Image classification is one of the core problems in computer vision, with a great number of practical applications included, such as object detection, analysis of remote sensing images, recognition of ultrasound liver images, etc. Generally, image classification aims to classify images according to their visual contents. The visual contents often contain both interested object and unrelated background. With the advent of big data times, research on image classification has been a hot spot and a kind of promising work. However, there exist many challenges in this area, including intra-class variance, scale and viewpoint variation, background clutter, etc., which bring negative effects to the performance of the current methods. Two-sides efforts, multi-view learning and metric learning, have been made to improve classification accuracy.

Multi-view learning optimizes multiple functions together, each function corresponds one view, to extract the information from diverse views as much as possible. In image processing, multiple feature subsets from different views are extracted with consideration of the relation in inter-views. Multi-view learning has been studied extensively with a large number of methods proposed, which can be classified into three groups: co-training, multiple kernel learning and subspace learning[1]. Co-training[2] method optimizes a mutual model on two distinct views alternately. Multiple kernel learning[3] assigns kernels to different views and combine the kernels effectively. Subspace learning[4, 5] seeks for a latent subspace, which is shared by different views under the assumption that the multiple views are generated from the subspace. Metric learning aims to learn a data-dependent metric $M$ to measure the distance(squared) between patterns(images) $x_1, x_2$ as $d_M(x_1, x_2) = (x_1 - x_2)^\top M(x_1 - x_2)$ instead of the traditional Euclidean distance where $M$ is an identity matrix. The desired metric should be a positive semidefinite one, which can be decomposed as $M = A^\top A$. The distance can be rewritten as $d_A(x_1, x_2) = (Ax_1 - Ax_2)^\top (Ax_1 - Ax_2)$. In fact, the distance $d_M$ is the Euclidean distance in the transformed space with the linear transformation $A$. A number of methods on metric learning have been proposed since 2002, with one of the earliest work emerged. The principle idea is minimizing the intra-class distance and maximizing inter-class distance under the new metric. The research on metric learning can be divided into two directions: global metric learning and local metric learning. The work from global view aims to find a metric to optimize and constrain all the rules on the entire dataset, representative methods contains MLSI[6], ITML[7], MCML[8]. But the local ones only implement the criterions on local neighborhood. NCA[9], LMNN[10] are classical models in local metric learning. A

*Corresponding author. E-mail: tyj@ucas.ac.cn

great many methods and surveys have been made on metric learning [11–16]. The efficiency of metric learning has been verified in improving the performance of $k$NN, $k$-means and other similarity search methods. Generally, image classification can be regarded as a similarity search problem where the measurement for similarity is critical. Then metric learning can be applied in image classification for better performance.

In this paper, we propose a novel framework for image classification. The technique of multi-view learning, metric learning and deep learning is fused together. For an image set with multiple-views, metric learning can be exploited to learn multiple metrics, each corresponds to a single view. Since metric learning is to learn a linear transformation essentially, which can be replaced by nonlinear transformation for better mapping ability. Deep neural networks(DNNs) are constructed to realize the goal of nonlinear mapping due to its strong ability in extracting features. The difference between inter-class distance and intra-class distance in every view and the correlation between different views of the same image are both maximized. So the discriminative information from different inputs and the view-specific information can be extracted as much as possible. Experiments results have validated the effectiveness of our new method.

The paper is organized as follows. Section 2 provides the details of model construction. The experiments are implemented in Section 3. Conclusions are given in Section 4.

## 2. PROPOSED FRAMEWORK

### 2.1. The model

Given a multi-view dataset with $m$ training examples from $c$ classes, $T = \{T_v \in R^{n_v \times m} \times Y\}_{v=1}^{V}$, where

$$T_v = \{(x_{v1}, y_1), (x_{v2}, y_2), \cdots, (x_{vm}, y_m)\} \tag{1}$$

is the feature set from $v$-th view and $y_i \in Y = \{1, 2, \cdots, c\}$ is the label corresponding to each feature input.

To make explicit nonlinear mapping, $V$ deep neural networks are constructed, each for a view. Suppose that the structure of all the networks are the same, but only different in the connected weights. Assume there are two indispensable layers, input layer and output layer, and $L$ hidden layers in each network with $d_l$ nodes in the $l$-th hidden layer. For each training input $x_{vi}$, its output of the first layer in the $v$-th network is $h_{vi}^1 = s(W_v^1 x_{vi} + b_v^1)$, where $W_v^1, b_v^1$ are the linear transformation and bias vector respectively, which will be learned by our model. And $s(\cdot)$ is a nonlinear active function. The output can be further simplified as $h_{vi}^1 = s(\hat{W}_v^1 \hat{x}_{vi})$, where $\hat{W}_v^1 = (W_v^1, b_v^1), \hat{x}_{vi} = (x_{vi}^\top, 1)^\top$. Similar as the traditional network, the output of the former layer is the input of the latter layer. So the output of the top hidden layer is $h_{vi}^L = s(W_v^L h_{vi}^{L-1} + b_v^L) = s(\hat{W}_v^L \hat{h}_{vi}^{L-1})$. Then the output $z_{vi} = s(\hat{W}_v^{L+1} \hat{h}_{vi}^L)$. Inspired by the idea of the local model in [17], a linear/nonlinear mapping that can make intra-class distance be smaller than inter-class distance on neighborhood level is enough for metric learning. For the $v-$th view, define two kinds of neighborhood for $z_{vi}$: intra-class neighborhood $S_{vi}$, which contains $K$ neighbors with the same label as $z_{vi}$, and inter-class neighborhood $D_{vi}$, which contains $K-1$ neighbors with different labels from $z_{vi}$. Then we construct the following optimization problem to realize our goal

$$\min_{\hat{W}} J_1 = \sum_{i=1}^{m} (d_1(z_{vi}) - C d_2(z_{vi})) + \lambda \sum_{l=1}^{L+1} \|\hat{W}_v^l\|^2 \tag{2}$$

where

$$d_1(z_{vi}) = \frac{1}{K} \sum_{z_{vk} \in S_{vi}} \|z_{vi} - z_{vk}\|^2 \tag{3}$$

$$d_2(z_{vi}) = \frac{1}{K-1} \sum_{z_{vk} \in D_{vi}} \|z_{vi} - z_{vk}\|^2 \tag{4}$$

denotes local intra-class distance and local inter-class distance of $z_{vi}$ respectively, with the neighborhood size $K$ and $K-1$ severally. The parameter $C$ is used to balance intra-class distance and inter-class distance. The last term in the problem (2) is used to avoid overfitting. Recent studies in multi-view learning have demonstrated that maximizing the correlations of different views can extract complementary information as much as possible[18]. So the differences between different views of the same pattern will be minimized in our model. For the $i-$th input, different views are enforced to be mapped into a single point, leading to the following problems

$$\min \qquad J_2 = \sum_{k,l=1}^{V} d(z_{ki}, z_{li}) \tag{5}$$

Then the framework of multi-view deep metric learning is established by adding up the above sub-problems

$$\begin{aligned} \min_{W,b} \quad & J = \sum_{v=1}^{V} \sum_{i=1}^{m} \alpha_v (d_1(z_{vi}) - C d_2(z_{vi})) + \frac{\mu}{2} \|\alpha\|^2 \\ & + \frac{\varepsilon}{2} \sum_{i=1}^{m} \sum_{k,l=1}^{V} d(z_{ki}, z_{li}) + \frac{\lambda}{2} \sum_{v=1}^{V} \sum_{l=1}^{L+1} \|\hat{W}_v^l\|^2 \quad (6) \\ s.t. \quad & e^\top \alpha = 1 \quad (7) \end{aligned}$$

where the trade-off $\varepsilon, \lambda, \mu$ are used to balance different terms and $e$ is a vector of ones with the length of $V$.

### 2.2. Alternative Optimization with gradient descent

To solve the optimization problem with respect to both linear transformation $\hat{W}$ and the weight $\alpha$, alternative optimization is used to obtain the solution alternately. First, the weight $\alpha$ is initialized and fixed, then the object function (6) is an unconstrained problem and gradient descent is adopted to solve

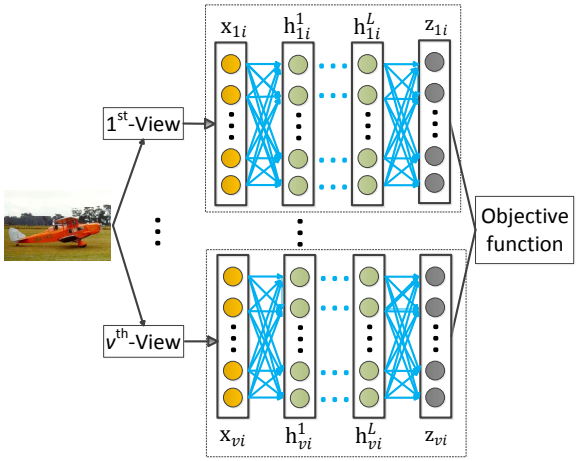



**Fig. 1**. The basic structure of our proposed model for image classification.

problem iteratively. The gradient of the objective function with respect to $\hat{W}_v^l$ is

$$\frac{\partial J}{\partial \hat{W}_v^l} = \alpha_v \sum_{i=1}^{m} \frac{\partial}{\partial \hat{W}_v^l}(d_1 - Cd_2) + \frac{\varepsilon}{2} \sum_{i=1}^{m} \sum_{l \neq v} \frac{\partial}{\partial \hat{W}_v^l} d(z_{ki}, z_{li}) + \lambda \hat{W}_v^l \quad (8)$$

For the output layer,

$$\frac{\partial d_1}{\partial \hat{W}_v^{L+1}} = \frac{2}{K} \sum_{z_{vk} \in S_{vi}} (z_{vi} - z_{vk}) \odot (z'_{vi} - z'_{vk}) \quad (9)$$

$$\frac{\partial d_2}{\partial \hat{W}_v^{L+1}} = \frac{2}{K-1} \sum_{z_{vk} \in D_{vi}} (z_{vi} - z_{vk}) \odot (z'_{vi} - z'_{vk}) \quad (10)$$

$$\frac{\partial}{\partial \hat{W}_v^{L+1}} d(z_{ki}, z_{li}) = 2(z_{vi} - z_{li}) \odot z'_{vi} \quad (11)$$

where$z'_{vi} = \sigma(z_{vi})(h^L_{vi})^\top$and $\sigma(a) = a \odot (\mathbf{1}_a - a)$, $a$ is a vector, $\mathbf{1}_a$ is a vector of ones with the same length as $a$. The operation $\odot$ denotes component-wise multiplication. For the $l-$th$(1 \leq l \leq L)$ hidden layer,

$$\frac{\partial d_1}{\partial \hat{W}_v^l} = (\hat{W}_v^{l+1})^\top \frac{\partial d_1}{\partial \hat{W}_v^{l+1}} \quad (12)$$

$$\frac{\partial d_2}{\partial \hat{W}_v^l} = (\hat{W}_v^{l+1})^\top \frac{\partial d_2}{\partial \hat{W}_v^{l+1}} \quad (13)$$

$$\frac{\partial}{\partial \hat{W}_v^l} d(z_{ki}, z_{li}) = (\hat{W}_v^{l+1})^\top \frac{\partial}{\partial \hat{W}_v^{l+1}} d(z_{ki}, z_{li}) \quad (14)$$

and $z'_{vi}$ will be changed as

$$z'_{vi} = \sigma(z_{vi}) \odot \sigma(h^L_{vi}) \odot \cdots \odot \sigma(h^l_{vi})(h^{l-1}_{vi})^\top \quad (15)$$

with the definition $h^0_{vi} = x_{vi}$. Then the linear transformation will be updated by

$$\hat{W}_v^l = \hat{W}_v^l - \eta \frac{\partial J}{\partial \hat{W}_v^l} \quad (16)$$

After obtaining the weight matrix $\hat{W}$, the following Lagrange problem is constructed

$$F = J - \gamma(e^\top \alpha - 1) \quad (17)$$

then $\alpha$ can be calculated based on the KKT condition,

$$\alpha = \frac{\mu e + e^\top \kappa e - V\kappa}{\mu V} \quad (18)$$

where $\kappa = (\kappa_1, \cdots, \kappa_V) \in R^V$ and $\kappa_v = \sum_{i=1}^{m}(d_1(z_{vi}) - Cd_2(z_{vi})), v = 1, \cdots, V$. The detailed procedure of MVDML is summarized in Algorithm 1.

**Algorithm 1** Multi-view deep metric learning for image classification(MVDML)

**Input:** The training set $T$; The penalty parameters $C, \varepsilon, \lambda, \mu$, gradient step-size $\eta$, maximum of iterations $T$.
**Output:** The target weights $\hat{W}_v^l, v = 1, \cdots, V, l = 1, \cdots, L+1$;
**Procedure:**
1. Let $t = 1$ and initialize $W_v^l$ as identity matrix;
2. Update $\hat{W}_v^l, v = 1, \cdots, V, l = 1, \cdots, L+1$ alternately using gradient ascent method by (16);
3. Update $\alpha$ by (18) and calculate the value of objective function (6) as $J_t$;
4. Let $t = t + 1$, if $t > T$ or $|J_t - J_{t-1}| < \delta (0 < \delta \ll 1)$, stop iteration and obtain the output, otherwise go to step 2.

### 2.3. Predict new images

Given a test image with $V$ views, all of its views will be input to corresponding networks learned from the training images. Suppose that the outputs are $z_1, z_2, \cdots, z_V$ and their nearest neighbors from the $c$-th class of the train set can be found, $z'_1, z'_2, \cdots, z'_V$. The distance between the test image and the nearest neighbor in $c$-th class is $d^c = \sum_{v=1}^{V} \alpha_v \|z_v - z'_v\|_2^2$. So the label of the test image is $y = \arg\min_c d^c$.

## 3. EXPERIMENTS

In the section, we will make numerical experiments to validate the effectiveness of our method in image classification. Three public and classical image datasets, **Caltech**, **Galaxy** and **GRAZ02**, were selected to make comparisons. The Caltech dataset contains 600 images from

six classes, including airplane, car, face, leave, motorbike and background. The dataset was collected by the student from California Institute of Technology. The Galaxy dataset(http://zoo1.galaxyzoo.org) contains 522 galaxy images with three kinds of shapes, edge-on, elliptical and spiral. The GRAZ02 dataset[19] consists of four classes of images, bike, car, person and background(Environment without bike, car or person). Three new image datasets were extended from GRAZ02, including **Bike**, **Car** and **Person**, each of which contains two classes, itself and background. Two common used feature extraction methods, HOG(Histogram of Oriented Gradient) feature[20] and LBP(Local binary pattern)[21], were used to construct two views for every image. Our method was compared with four single-view methods, including $k$NN with Euclidean distance, MCML, LMNN and ITML. For our model, the neural networks were all designed with three layers, input layer, one hidden layer and output layer. There are 500, 100 and 50 neural nodes in the three layers respectively. The penalty parameters $C, \lambda, \varepsilon, \mu$ were all set to be 1. The learning rate $\eta$ was set as 0.01. Three-fold crossvalidation method was used to select the best parameters for each model. The activate function is sigmoid function.

**Table 1**. Classification error rates on single view and multi-view

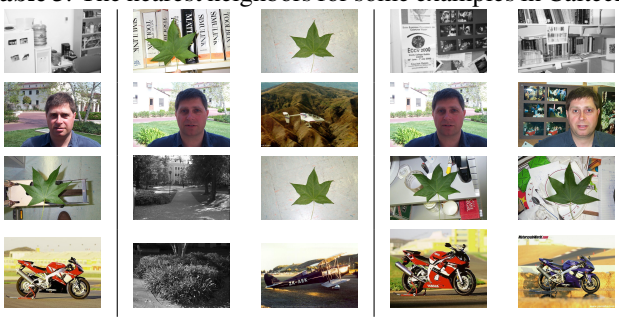
| Datasets | View | Euc | MCML | LMNN | ITML | MVDML |
|---|---|---|---|---|---|---|
| Caltech (600&6) | Single | 11.3±5.0 | 7.7±0.3 | 8.0±2.2 | 7.8±2.9 | \ |
| | | 18.2±0.8 | 15.3±2.8 | 15.0±2.0 | 11.8±0.6 | \ |
| | Multiple | 11.3±5.0 | 7.0±1.5 | 7.5±1.7 | 6.3±1.8 | **6.3**±0.3 |
| Galaxy (522&3) | Single | 19.2±2.6 | 14.0±4.4 | 14.2±3.4 | 14.0±3.4 | \ |
| | | 20.5±3.4 | 20.5±3.2 | 21.1±0.9 | 15.7±1.3 | \ |
| | Multiple | 19.2±2.3 | 13.2±3.0 | 14.2±2.9 | 14.0±4.4 | **11.7**±0.3 |
| GRAZ02 (800&4) | Single | 58.2±3.0 | 55.3±2.0 | 53.6±2.7 | 57.9±2.6 | \ |
| | | 51.3±2.5 | 50.1±4.1 | 38.3±2.1 | 52.5±5.8 | \ |
| | Multiple | 57.7±3.2 | 52.5±4.8 | 48.3±0.6 | 58.9±3.7 | **42.8**±1.5 |
| bike (745&2) | Single | 40.1±2.6 | 30.6±1.9 | 38.8±3.7 | 36.8±1.5 | \ |
| | | 31.6±5.4 | 32.4±1.2 | 31.3±2.7 | 31.6±0.8 | \ |
| | Multiple | 40.1±2.4 | **30.2**±2.4 | 31.7±1.7 | 35.8±2.0 | 32.8±2.9 |
| car (800&2) | Single | 43.0±4.9 | 39.5±1.5 | 42.8±1.8 | 41.2±4.0 | \ |
| | | 40.4±0.4 | 39.3±1.1 | 35.9±3.1 | 36.7±1.7 | \ |
| | Multiple | 42.6±4.9 | 37.7±1.5 | **36.5**±1.9 | 40.7±4.1 | 38.0±2.1 |
| person (691&2) | Single | 36.9±4.5 | 25.5±1.5 | 30.3±0.9 | 31.2±5.8 | \ |
| | | 35.9±3.1 | 35.2±1.2 | 32.1±1.1 | 34.2±3.6 | \ |
| | Multiple | 36.8±4.5 | **27.2**±1.8 | 28.8±2.0 | 30.7±5.1 | 28.4±2.0 |

The average error rates of each method on the six datasets are shown in Table 1. For each dataset, the first line and second line denotes the results on HOG and LBP feature respectively. For the baseline methods, the two features were combined into a long feature vector and the corresponding results are shown on the third line. The lowest error rates of multi-view are in boldface type. It can be seen that our method performs best on three datasets and second on another datasets. For all the selected datasets, our model obtains competitive performance with the baseline models. We compared the CPU time of different models on three datasets and the results are provided in Table 2. For MCML, LMNN and ITML, the former two times denotes HOG and LBP respectively and the third denotes the time on multi-view situation. It is obviously that our approach runs much faster than the three previous models. We selected four query images from the Caltech dataset and provide their nearest neighbors in $k$NN with Euclidean distance and MVDML. The results are given in Tabel 3. For each query image in the left column, its nearest neighbors search by Euclidean distance from HOG and LBP representation are the second and third images in the same row. The last two images are the nearest pair found by MVDML.

**Table 2**. CPU time of different methods(seconds).

| Datasets | MCML | LMNN | ITML | MVDML |
|---|---|---|---|---|
| Caltech | 605+685/2056 | 313+368/533 | 119+97/154 | 49s |
| Galaxy | 427+375/1783 | 93+357/172 | 119+113/133 | 44s |
| GRAZ02 | 1032+929/4210 | 82+474/155 | 121+107/172 | 67s |

**Table 3**. The nearest neighbors for some examples in Caltech.



## 4. CONCLUSIONS

In this paper, a novel framework is proposed to improve the performance of $k$NN on image classification, with the technique of multi-view learning, deep learning and metric learning embedded. For each image with multiple views, it is important but hard to extract complementary information as much as possible and combine the information in a harmonious way. Multiple deep neural networks are constructed, each corresponds to a single view, to make nonlinear transformation for the inputs. In output space, the model aims to maximize the difference between intra-class distance and inter-class distance. Also the difference between different views of the same image is minimized to achieve the goal of maximization of the correlation between distinct views. Experiments on benchmark datasets demonstrate that our method is effective in classifying images with multi-views in much less time than compared approaches.


# Acknowledgement

This work has been partially supported by grants from National Natural Science Foundation of China (Nos .61472390, 11271361, 71331005, and 11226089), Major International (Regional) Joint Research Project (No. 71110107026) and the Beijing Natural Science Foundation (No.1162005).

## REFERENCES


[1] Chang Xu, Dacheng Tao, and Chao Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[2] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[3] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine learning research*, vol. 5, no. Jan, pp. 27–72, 2004.

[4] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[5] Shotaro Akaho, "A kernel method for canonical correlation analysis," *arXiv preprint cs/0609071*, 2006.

[6] Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Y Ng, "Distance metric learning with application to clustering with side-information," in *Advances in neural information processing systems*, 2002, pp. 505–512.

[7] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 209–216.

[8] Amir Globerson and Sam T Roweis, "Metric learning by collapsing classes," in *Advances in neural information processing systems*, 2005, pp. 451–458.

[9] Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Ruslan Salakhutdinov, "Neighbourhood components analysis," in *Advances in neural information processing systems*, 2004, pp. 513–520.

[10] Kilian Q Weinberger and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[11] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka, "Distance-based image classification: Generalizing to new classes at near-zero cost," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 11, pp. 2624–2637, 2013.

[12] Yiming Ying and Peng Li, "Distance metric learning with eigenvalue optimization," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1–26, 2012.

[13] Chunhua Shen, Junae Kim, Lei Wang, and Anton Hengel, "Positive semidefinite metric learning with boosting," in *Advances in neural information processing systems*, 2009, pp. 1651–1659.

[14] Liu Yang and Rong Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, 2006.

[15] Brian Kulis, "Metric learning: A survey," *Foundations and Trends in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2012.

[16] Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou, "Multi-manifold deep metric learning for image set classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1137–1145.

[17] Dewei Li and Yingjie Tian, "Global and local metric learning via eigenvectors," *Knowledge-Based Systems*, 2016.

[18] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2160–2167.

[19] Andreas Opelt and Axel Pinz, "Object localization with boosting and weak supervision for generic object recognition," in *Scandinavian Conference on Image Analysis*. Springer, 2005, pp. 862–871.

[20] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, vol. 1, pp. 886–893.

[21] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen, "Face recognition with local binary patterns," in *European conference on computer vision*. Springer, 2004, pp. 469–481.