15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Supervised Multi-modal Dictionary Learning for Clothing Representation

Qilu Zhao Jiayan Wang Zongmin Li
China University of Petroleum (East China)
No. 66 , Changjiang West Road , Huangdao District , Qingdao , China
{734727745, 867177665, 403042638}@qq.com

Abstract

Clothing appearances have complex visual properties, such as color, texture, shape and structure. Different modalities of visual features provide information complementary to each other. Combining multi-modal visual features can lead to a comprehensive description of Clothing appearances. Meanwhile, categories provide sufficient semantic information, which can lead to discriminative representations. Clothing categories exhibit hierarchical structure, which could benefit the learning altorithm. In this paper, we propose a multi-view learning algorithm, named Supervised Multi-modal Dictionary Learning (SMMDL), which learns a latent space encoding multi-modal visual properties and semantic relationships between clothing samples. Experiments on the image classification task show that SMMDL outperforms baseline methods.

1 Introduction

Clothing appearances have complex visual properties, such as color, texture, shape and structure. Using one single modality of visual features is incapable of representing large-scale clothing images effectively. A straightforward solution for combining multi-modal visual features is to concatenate the feature vectors to form a new vector. However, this method ignores the specific statistical properties of different modalities and usually incurs the curse of dimensionality problem.

Combing multi-modal visual features can be solved by multi-view learning approach, which aims at learning unified representations from multi-view data. A growing area in the multi-view learning literature is multi-view latent subspace learning, which aims at obtaining a unified latent subspace shared by multiple views. This approach includes Canonical Correlation Analysis (CCA) [1], the shared Kernel Information Embedding model (sKIE) [3], and the shared Gaussian Process Latent Variable Model (shared GPLVM) [4,5,6]. However, these methods do not account for the independent parts of the views, and therefore either totally fail to represent them, or mix them with the information shared by all views.

In our opinion, useful information patterns hidden in multi-view data might not be associated with all the views. For example, images of concept plaid shirts exhibit visual patterns which are associated with texture features but independent on color features, since the plaid can be in different colors. Thus we need to design a latent space learning algorithm which can correctly factorize the information of all views into shared and private parts. Fortunately, to the best of our knowledge there are already three papers trying to solve this problem [7,8,9]. We chose to use Jia Yangqings work [9], because it avoids the computational burden of other techniques [7,8].

In supervised tasks, its important to exploit the label information. In dictionary learning, some works [10,11,12,13] introduce laplacian regularization terms or loss functions into the optimization problem. In [10,11], a laplacian regularization term was introduced to preserve the consistence of sparse codes for the similar local features. In [12,13], loss functions were incorporated into the objective function, which enhances the discriminative power of sparse codes. However, incorporating loss function also makes the learning process more complex and brings more computational burden. Besides, all the works mentioned above haven't considered the semantic hierarchy of class labels.

Clothing categories exhibit hierarchical structure, which encodes sufficient semantic relationships between clothing samples. For example, jacket and shirt belong to tops, while trouser and breeches belong to bottoms. The latent space learning process could benefit from the hierarchical structure, which exploits the inter-class semantic correlations in the label space. The idea of exploiting category hierarchy has been extensively studied in the computer vision community under different frameworks. For example, in [14] Bengio has introduced an approach for fast multi-class classification by learning label embedding trees; in [15] Binder has proposed a structured learning framework to study the problem of classifying images into a predetermined taxonomy (category hierarchy).

If a feature assigns similar values to the instances that are close to each other on a given graph, it indicates that the feature is consistent with the graph structure. The laplacian regularization term [10,11] quantifies how much the feature varies locally or how smooth it is over the Graph. Thus, using the label information to construct the graph, the laplacian regularization term can quantify how consistent the feature is with the label information. Inspired by this, in this paper we leverage the laplacian regularization term used in [10,11] to capture the semantic relationships between clothing samples in a hierarchical way, which hasn't been exploited in the previous works.

2 Supervised Multi-modal Dictionary Learning

In this section, we introduce Supervised Multi-modal Dictionary Learning. From the set of clothing images, we first extract four modalities of visual features: color, texture, shape and structure. For each image, 7936D Histogram of Oriented Gradient [16] is extracted to characterize its textural property. 10D

Fourier Descriptor [17] is extracted to characterize its shape property. 512D GIST [18] is extracted to characterize its structure property. And 512D color histogram [19] (i.e., R, G, B color channels are first quantized into 8 units) is extracted to characterize its color property. Let $\{X^{(v)}\}_{v=1}^{4}$ denotes the feature matrix of N images, where $X^{(v)} \in R_{+}^{M_v \times N}$ contains the feature vectors for the v^{th} modality. We aim to find an embedding $\alpha \in R_{+}^{K \times N}$ of the data into a K-dimensional latent space and a set of dictionaries $\{D^{(v)}\}_{v=1}^{4}$, with $D^{(v)} \in R_{+}^{M_v \times K}$ the dictionary entries for the v^{th} modality.

2.1 Multi-modal Dictionary Learning

The basic optimization framework of SMMDL tries to learn a common latent space via a shared latent embedding α:

$$\min_{\{D^{(v)}\}_{v=1}^{H},\alpha} \frac{1}{2}\sum_{v=1}^{4}\|X^{(v)} - D^{(v)}\alpha\|_F^2 \quad s.t. D_{ik}^{(v)} \geq 0, 1 \geq \alpha_{kj} \geq 0, \forall i,j,k,v \tag{1}$$

in this way, each image is forced to have the same embedding under different modalities, and the dictionaries of different modalities are coupled together through α. Furthermore, as explained in Section 1, we aim to find a latent space that naturally separates the information shared among several modalities from the information private to each modality. We leverage Jia's work [9] to enforce structured sparsity on the dictionary entries, which lets latent dimensions be shared across any subset of the modalities rather than across all modalities only. For each $D^{(v)}$, a structured sparseness regularizer is added to the objective function (1) to encourage some columns to be zeroed-out. One can achieve structured sparsity via $L_{1,q}$ norm where q is an integer ranging from 1 to ∞ :

$$\|D^{(v)}\|_{1,q} = \sum_{k=1}^{K}\|D_k^{(v)}\|_q \tag{2}$$

where $D_k^{(v)}$ denotes the kth column of $D^{(v)}$. In practice, we chose the $L_{1,\infty}$ norm regularizer which has proven more effective than the others [24]:

$$\|D^{(v)}\|_{1,\infty} = \sum_{k=1}^{K}\max_{1\leq i\leq M_v}|D_{ik}^{(v)}| \tag{3}$$

now, we can re-formulating (1) as:

$$\min_{\{D^{(v)}\}_{v=1}^{H},\alpha} \frac{1}{2}\sum_{v=1}^{4}\|X^{(v)} - D^{(v)}\alpha\|_F^2 + \gamma\sum_{v=1}^{4}\|D^{(v)}\|_{1,\infty} \quad s.t. D_{ik}^{(v)} \geq 0, 1 \geq \alpha_{kj} \geq 0, \forall i,j,k,v \tag{4}$$

2.2 Hierarchical Laplacian regularization term

In [10], the laplacian regularization term is defined as follows:

$$\sum_{ij}\|\alpha_i - \alpha_j\|^2 W_{ij} \tag{5}$$

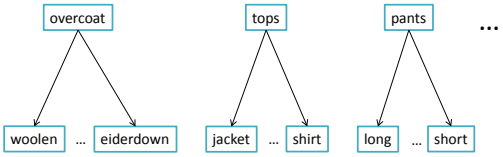

Figure 1. An example, a two-level hierarchical structure of clothing categories

where α_i denotes the i-th column of α, W denotes the similarity matrix of a graph G. The function above quantifies how much α varies locally or how smooth it is over G. A smooth α assigns similar vectors to the instances that are close to each other on G, thus it is consistent with the graph structure. This observation is the motivation behind [10].

Clothing categories exhibit hierarchical structure. As shown in Figure 1, we leverage a two-level hierarchy to guide the latent space learning. For each level, we construct a graph, where the nodes are training images and each edge is weighted by the similarity between the nodes it's associated with. Using the label information to compute the similarity, the laplacian regularization term can measure the latent space's consistency with the class labels of each level.

Using the class labels of a certain level, the similarity between images can be defined by:

$$W_{ij} = \begin{cases} \frac{1}{N_l} - \frac{1}{N}, y_i = y_j = l \\ 0, otherwise \end{cases} \tag{6}$$

where N_l denotes the number of instances in class l, y_i denotes the class label of training image i, n is the number of training images.

We use G_1 to denote the graph of top level in Figure 2, and G_2 denotes the graph of bottom level. W^1 denotes the similarity matrix of graph G_1, W^2 denotes the similarity matrix of graph G_2. The hierarchical Laplacian regularization term is defined as follows:

$$\begin{aligned} &\frac{\eta}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}W_{ij}^1\|\alpha_i - \alpha_j\|_2^2 + \frac{\xi}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}W_{ij}^2\|\alpha_i - \alpha_j\|_2^2 \\ &= \frac{\eta}{2}tr[\alpha L^1(\alpha)^T] + \frac{\xi}{2}tr[\alpha L^2(\alpha)^T] \end{aligned} \tag{7}$$

where $tr(.)$ denotes the trace of a matrix, $L^1 = U^1 - W^1$ is the Laplacian matrix for graph G^1 with the (i,i)th element of the diagonal matrix U^1 equals $\sum_{j=1}^{N}W_{ij}^1$ (L^2 is for G^2).

By synthesizing the above objectives, the optimization problem of SMMDL is formulated as:

$$\begin{aligned} &\min_{\{D^{(v)}\}_{v=1}^{H},\alpha} \frac{1}{2}\sum_{v=1}^{4}\|X^{(v)} - D^{(v)}\alpha\|_F^2 + \gamma\sum_{v=1}^{4}\|D^{(v)}\|_{1,\infty} \\ &+ \frac{\eta}{2}tr[\alpha L^1(\alpha)^T] + \frac{\xi}{2}tr[\alpha L^2(\alpha)^T] \\ &s.t. D_{ik}^{(v)} \geq 0, 1 \geq \alpha_{kj} \geq 0, \forall i,j,k,v \end{aligned} \tag{8}$$

2.3 Optimization

we develop an iterative algorithm to optimize the variables $\{D^{(v)}\}_{v=1}^{4}$ and α alternatively.

It is clear that, when α is fixed, $\{D^{(v)}\}_{v=1}^4$ are independent with one another. Since the optimization method is the same, here we just focus on an arbitrary modality and use X and D to denote respectively the feature matrix and the dictionary for the modality. The subproblem involving D can be written as:

$$\min_D \frac{1}{2}\|X - D\alpha\|_F^2 + \gamma\|D\|_{1,\infty}, s.t. D_{ik} \geq 0, \forall i,k \quad (9)$$

To optimize function (9), we use an optimization algorithm developed in [21], which is based on the composite gradient mapping technique proposed for minimizing composite objective functions [22]. The idea is to iteratively minimize an auxiliary function and adjust the guess of the Lipschitz constant of the first term of function (9), so that the objective function can be decreased as fast as possible. The details of the Composite Gradient Mapping algorithm can be found in [21].

For the optimization of α, we leverage a multiplicative update algorithm [21], which is based on a general multiplicative optimization scheme proposed in [23]. The update rule is as follow:

$$\begin{aligned}
\alpha_{kj}^{t+1} &= \min\{1, \frac{-B_{kj} + \sqrt{B_{kj}^2 + 4A_{kj}C_{kj}}}{2A_{kj}}\alpha_{kj}^t\} \\
A_{kj} &= (P\alpha_j^t)_k + \eta((U^1 + \frac{\xi}{\eta}U^2)\alpha_k^t)_j, \\
B_{kj} &= -Q_{kj}, \\
C_{kj} &= \eta((W^1 + \frac{\xi}{\eta}W^2)\alpha_k^t)_j
\end{aligned} \quad (10)$$

where $P = \sum_{v=1}^4 (D^{(v)})^T D^{(v)}$, $Q = \sum_{v=1}^4 (D^{(v)})^T X^{(v)}$. The detailed derivations are similar with those in [21], which can be found in the appendix of [21]. Given a new observation, the corresponding α^* can be obtained by solving the objective function (1).

3 Experiment

In order to evaluate the SMMDL, we collect two datasets from the Internet: one is collected from e-Commerce websites (e-commerce), such as JD.com and Amazon.com; the other is collected from bing.com and google.com (street), where we input keywords like "jeans street style" and download the search results. The first dataset includes 11,700 images with pure background, and the other includes 6,628 images with complex background. Both exhibit a two-level hierarchy, such as tops (shirts, jacket, sweater, ...), pants (jeans, chinos, cargo pants, ...) and skirt (wedding dress, cheongsam, formal dress, ...). Four fifths of images are randomly selected as training set, and the rest as test set. For classification, we use a kNN classifier(k=10) and the accuracy of the classifier on the test set is calculated.

Baseline methods include:

1 best view (Single-B): This baseline applies each view to the classification task,and reports the best performance.

2 feature concatenation (Concat): This method concatenates feature vectors of different modalities to form a united representation.

Table 1. Classification Performances

e-commerce	Accuracy	*street*	Accuracy
Single-B	0.4466	Single-B	0.2601
Concat	0.4517	Concat	0.2678
CCA	0.4796	CCA	0.2787
GMA	0.4973	GMA	0.3272
shared	0.48	shared	0.2873
SMMDL	**0.5520**	SMMDL	**0.3583**

3 shared multi-modal dictionary learning (shared): objective function (1) is used to learn the shared latent space.

4 Canonical Correlation Analysis (CCA): CCA [1] has been the workhorse for learning a common latent space which is evident from its wide-spread use in vision [2].

5 Generalized Multiview Analysis (GMA): GMA [2] is a supervised extension of Canonical Correlational Analysis (CCA).

Tables 1 shows the classification performance results on two datasets. Observations are as follows. First, methods that made use of multi-modal features outperformed Single-B, which only used one modality. Second, supervised methods outperformed unsupervised methods, which indicated that exploiting label information could lead to more discriminative latent features. Third, SMMDL outperformed other methods evaluated in this paper.

Figure 2 shows the influence of different parameter settings on the performance of SMMDL. We vary one parameter at a time while fixing the other two. The general behavior of the three parameters was the same: when increasing the parameter from 0, the performance curves first went up and then went down. This indicates that assigning moderate weights is important in the practice. The performance of SMMDL is sensitive to the value of γ. When the value of γ is larger than 80, the performance might drop drastically. For η and ξ, small values are better choices. We tested $\eta = 10$ and $\xi = 0$, which is not shown in Figure 2, and the accuracy of SMMDL on *e-commerce* dropped to 0.3356. This indicates that large value of η and ξ would lead to serious over-fitting problem.

4 Conclusions

In this paper, we propose a multi-view learning algorithm, named Supervised Multi-modal Dictionary Learning (SMMDL), which exploits multi-modal visual features and label information to learn a latent space for clothing representation. Experiments show that SMMDL has achieved competitive results.

References

[1] Hotelling H. Relations between two sets of variates[M] Breakthroughs in Statistics. Springer New York, 1992: 162-190.

[2] Sharma A, Kumar A, Daume H, et al. Generalized Multiview Analysis: A discriminative latent space[C]. Computer Vision and Pattern Recognition, 2012.

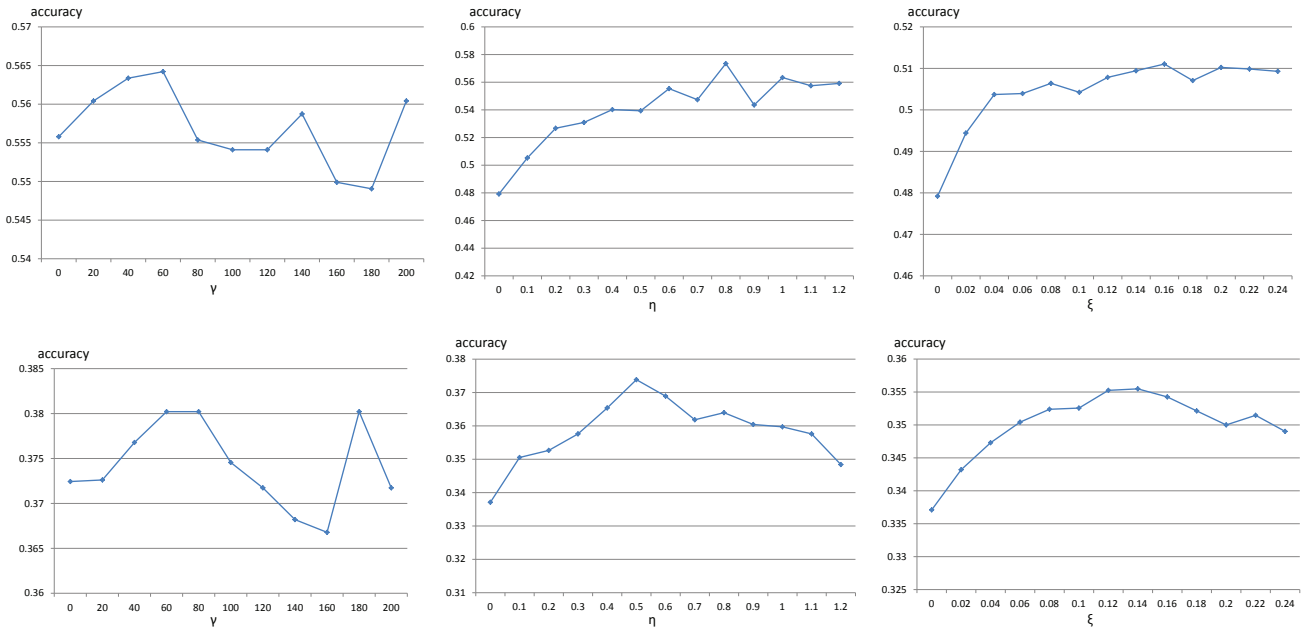

Figure 2. Influence of different parameter settings on the performance of SMMDL. From top to bottom: e-commerce, street. From left to right: varying γ while setting $\eta = 1$ and $\xi = 0$ (e-commerce) and $\eta = 0.5$ and $\xi = 0$ (street), and varying η while setting $\gamma = 200$ and $\xi = 0$ (e-commerce) and $\gamma = 100$ and $\xi = 0$ (street), and varying ξ while setting $\gamma = 200$ and $\eta = 0$ (e-commerce) and $\gamma = 100$ and $\eta = 0$ (street).

[3] L. Sigal, R. Memisevic, and D. J. Fleet. Shared kernel information embedding for discriminative inference.In Conference on Computer Vision and Pattern Recognition, 2009.

[4] A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao. Learning shared latent structure for image synthesis and robotic imitation. In Neural Information Processing Systems, pages 1233C1240, 2006.

[5] C. H. Ek, P. Torr, and N. Lawrence. Gaussian process latent variable models for human pose estimation.In Joint Workshop on Machine Learning and Multimodal Interaction, 2007.

[6] R. Navaratnam, A. Fitzgibbon, and R. Cipolla. The Joint Manifold Model for Semi-supervised Multivalued Regression. In International Conference on Computer Vision, Rio, Brazil, October 2007.

[7] M. Salzmann, C.-H. Ek, R. Urtasun, and T. Darrell. Factorized orthogonal latent spaces. In International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, May 2010.

[8] G. Leen. Context assisted information extraction. PhD thesis, University the of West of Scotland, University of the West of Scotland, High Street, Paisley PA1 2BE, Scotland, 2008.

[9] Jia Y., Salzmann M.and Darrell T. Factorized latent spaces with structured sparsity. Advances in Neural Information Processing Systems 23:982C990, 2010.

[10] Gao S, Tsang I W, Chia L, et al. Local features are not lonely C Laplacian sparse coding for image classification[C]. CVPR, 2010.

[11] Gao S, Tsang I W, Chia L, et al. Laplacian Sparse Coding, Hypergraph Laplacian Sparse Coding, and Applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 92-104.

[12] Yang J, Yu K, Huang T S, et al. Supervised translation-invariant sparse coding[C]. Computer Vision and Pattern Recognition, 2010.

[13] Mairal J, Bach F, Ponce J, et al. Supervised dictionary learning[C]. Neural Information Processing Systems, 2008.

[14] Bengio S, Weston J, Grangier D, et al. Label Embedding Trees for Large Multi-Class Tasks.[C]. Neural Information Processing Systems, 2010.

[15] Binder A, Muller K, Kawanabe M, et al. On Taxonomies for Multi-class Image Categorization[J]. International Journal of Computer Vision, 2012, 99(3): 281-301.

[16] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. Computer Vision and Pattern Recognition, 2005.

[17] D.S. Zhang, G. Lu, A comparative study of Fourier descriptors for shape representation and retrieval, in:Proceedings of the Fifth Asian Conference on Computer Vision (ACCV02).

[18] Oliva A, Torralba A. Building the gist of a scene: the role of global image features in recognition.[J]. Progress in Brain Research, 2006.

[19] M. J. Swain and D. H. Ballard, Color indexing, International Journal Computer Vision,vol. 7, no. 1, pp. 11C32, 1991.

[20] Fan R. K. Chung, Spectral Graph Theory, Regional Conference Series in Mathematics, number 92, 1997.

[21] Guan Z, Zhang L, Peng J, et al. Multi-View Concept Learning for Data Representation[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(11): 3016-3028.

[22] Nesterov Y. Gradient methods for minimizing composite functions[J]. Mathematical Programming, 2013, 140(1): 125-161.

[23] F. Sha, Y. Lin, L. K. Saul, and D. D. Lee, Multiplicative updates for nonnegative quadratic programming, Neural Comput., vol. 19, no. 8, pp. 2004C2031, 2007.

[24] Quattoni A, Carreras X, Collins M J, et al. An efficient projection for l1 infinity regularization.[C]. ICML, 2009.