PARTIALLY TAGGED IMAGE CLUSTERING

*Qiyue Yin, Shu Wu, Liang Wang**

Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{qyyin, shu.wu, wangliang}@nlpr.ia.ac.cn

ABSTRACT

With the growth of tagged images, researchers are using this highly semantic tag information to assist some vision tasks such as image clustering. However, users may not tag some images at all or some of the images are partially annotated, and this will lead to performance degradation, which is rarely considered by previous works. To alleviate this problem, we propose a new model for image clustering assisted by partially observed tags. Our model enforces sparse representations obtained through sparse coding and latent tag representations learned via matrix factorization to be consistent with the partial image-tag observations. The partition of image database is finally performed using clustering algorithms (e.g., k-means) on the sparse representations. Extensive experiments demonstrate that the proposed model performs better than the state-of-the-art methods.

Index Terms— Image clustering, partially observed tag information, multi-view clustering, sparse coding

1. INTRODUCTION

Image clustering, which assigns images into different groups, plays an important role for image organization and visualization [1]. Traditional image clustering algorithms usually resort to visual features such as the SIFT descriptor. However, using such low-level visual features is always ineffective because of the problem of the semantic gap [2, 3]. Researchers are now exploring the textual information surrounding the images, such as the tags, as complementary high-level semantic information to boost the clustering performance.

Several works have been proposed fusing visual and textual features to improve clustering [4, 5, 6, 7, 8]. Cai et al. [4] proposed a hierarchical clustering model to fuse the visual, textual and link information for clustering of Web image search results. Similarly, Peng et al [9] utilized tags to obtain topics as the first clustering layer and then used the visual features for more sophisticated clusters. Furthermore, Rege [8] proposed a co-clustering based framework for simultaneously integrating visual and textual features and then graph theory was applied for the final clustering. Recently, multi-view clustering, which fuses multiple sources of information for clustering tasks, provides a natural way for combining the visual and textual features. It achieves nice results and draws significant attention nowadays [10, 11, 12, 13, 14]. Generally, a wide variety of multi-view clustering works aim at finding a low dimensional embedding among multiple features, and the complementary information is expected to be maximized in this learned latent space. Several typical examples such as [15, 16, 11, 17, 18] obtain promising clustering results.

Although some works have been proposed for image clustering utilizing both visual and textual features, few of them consider the scenario that the textual information is incomplete, which commonly exists in real applications. Compared with the visual features that can be extracted by representative descriptors, images sometimes may not be annotated or only given a few tags that are not abundant for image description. In this circumstance, conventional methods may face the problem of performance degradation due to the great dependency on the complete textual information. It should be noted that several works [17, 18] have been proposed to solve the incomplete view problem for multi-view clustering. However, these methods mainly focus on the text data (e.g., Web pages clustering) and may not be suitable for image clustering.

In this paper, we propose a novel model that focuses on image clustering assisted by partially observed tag information. Our model utilizes tags to assist learning of the visual representations, which consists of two parts. The first part is sparse coding. We learn sparse representations based on visual features, which can capture salient structures of the images [2]. In the second part, we learn the latent representation for each tag, and keep the sparse representations and the tag representations being consistent with the partial image-tag observations. Furthermore, an importance matrix is employed to deal with the situation that a tag is related to an image but not be observed. Finally, image clustering is achieved by performing clustering algorithms (e.g., k-means) on the learned sparse representations.

Our contribution in this paper is summarized as follows: 1) A novel model for image clustering assisted by partially

*The corresponding author

observed tag information is proposed, which is designed for the scenario that the tags for some images are totally missing or partially observed. To the best of our knowledge, this scenario is rarely considered for image clustering. 2) An effective optimizing algorithm for the proposed model is developed. And extensive experiments on three real world datasets show that the proposed model obtains better clustering results compared with several state-of-the-art methods.

2. MODEL

2.1. Problem overview

We use X^T to represent the transpose of a matrix X. X^i and X_i indicate the i-th row and the i-th column of X respectively. X_{ij} is the entity of i-th row and j-th column of X. For two matrices with the same size, we use $\odot$ to denote the element-wise product. Suppose we have n images and its visual feature is denoted as $X \in \Re^{p\times n}$ with p as the dimensionality. As for the textual feature T, assume that we have q tags and if the i-th tag is annotated to the j-th image, T_{ij} is assigned to be 1, otherwise 0. It should be noted that T can be incomplete, which means some of the images have no tags or some tags of an image may be missing. Our task is to cluster such partially tagged images.

2.2. Formulation

Inspired by sparse coding, we assume that an image can be represented as a spare linear combination of the learned dictionary. Furthermore, we learn the latent representation for each tag and use this latent feature to assist the learning of sparse coding. Thus, the objective we are going to optimize is listed as follows:

$$\begin{aligned}\min_{B,S,C}\ & ||X-BS||_F^2+\alpha||S||_1 \\ & +\lambda(\textstyle\sum_{ij\in O}\left(T_{ij}-C^iS_j\right)^2+\beta||C||_F^2) \\ s.t.\quad & ||B_t||^2\le 1,\ \forall t\end{aligned} \tag{1}$$

where B and S are the learned dictionary and the sparse representation respectively. C is the latent representation for all the tags and O is the observed tag-image set. The parameters α, λ and β are scalars balancing different terms. The constraints on B is usually adopted by sparse coding as in [19]. After the optimization of (1), we can use the clustering algorithms, such as k-means, on S for final data partitioning.

The purpose of the third term in (1) is to enforce the learned two representations to be consistent with the partial image-tag observations. More specifically, we model the consistency using the latent factor model via matrix factorization, namely, we constrain the dot product of the learned tag representation and the sparse representation to approach the matrix T. Intuitively, C^iS_j can be viewed as a linear sum that represents how the i-th tag is related to the j-th image. By doing so, the sparse representations of two images will be close if they have similar tag information. The term $\|C\|_F^2$ is a regularizer to avoid over-fitting.

In our hypothesis, $T_{ij}=0$ can be interpreted as either the i-th tag is not related to the image j or it is missing. So we employ an importance matrix $I \in \Re^{q\times n}$ with the same size of T to alleviate the missing situation. Similar to [20], I_{ij} is assigned to be a small value when $T_{ij}=0$. And the objective is reformulated as:

$$\begin{aligned}\min_{B,S,C}\ & ||X-BS||_F^2+\alpha||S||_1 \\ & +\lambda(\textstyle\sum_{ij} I_{ij}(T_{ij}-C^iS_j)^2+\beta||C||_F^2) \\ s.t.\quad & ||B_t||^2\le 1,\ \forall t\end{aligned} \tag{2}$$

where I is defined as follows:

$$I_{ij}=\begin{cases} a & if\ T_{ij}=1 \\ b & if\ T_{ij}=0\end{cases} \tag{3}$$

where a and b are two scalars satisfying $a>b>0$.

Finally, we write the above equation as a compact matrix form:

$$\begin{aligned}\min_{B,S,C}\ & ||X-BS||_F^2+\alpha||S||_1 \\ & +\lambda(||L\odot(T-CS)||_F^2+\beta||C||_F^2) \\ s.t.\quad & ||B_t||^2\le 1,\ \forall t\end{aligned} \tag{4}$$

where $L=I^{1/2}$ is the element-wise square root of matrix I.

3. SOLUTION

Since the variables B, S and C are coupled together and it may be difficult to solve them jointly, we propose to optimize the three variables alternatively until convergence. Note that the convergence analysis is similar to that of sparse coding [19] and is omitted here due to space limitation.

Solve S with B and C fixed. The problem becomes:

$$\min_S \|X-BS\|_F^2+\lambda\|L\odot(T-CS)\|_F^2+\alpha\|S\|_1 \tag{5}$$

For each column S_i, we have:

$$\min_{S_i}\left\|\begin{bmatrix}X_i\\ \sqrt{\lambda}L_i\odot T_i\end{bmatrix}-\begin{bmatrix}B\\ \sqrt{\lambda}diag\left(L_i\right)C\end{bmatrix}S_i\right\|^2+\alpha\|S_i\|_1 \tag{6}$$

where $diag(v)$ denotes a diagonal matrix with its diagonal elements being the vector v. This is a standard sparse representation problem, which can be solved using SLEP packages[1].

Solve C with B and S fixed. The problem is written as:

$$\min_C \|L\odot(T-CS)\|_F^2+\beta\|C\|_F^2 \tag{7}$$

For each row C^i, the problem is simplified as:

$$\min_{C^i}\left\|T^idiag(L^i)-C^iSdiag(L^i)\right\|^2+\beta\left\|C^i\right\|^2 \tag{8}$$

where we can easily obtain the analytic solution for C^i.

[1] http://parnec.nuaa.edu.cn/jliu/largeScaleSparseLearning.htm

Solve B with S and C fixed. We have the following problem:

$$\min_{B} \|X - BS\|_F^2 \qquad s.t. \|B_t\|^2 \leq 1, \forall t \tag{9}$$

which can be optimized through the Lagrangian method. Suppose the size of the dictionary is k, then it becomes:

$$L(B, \phi) = \|X - BS\|_F^2 + \sum_{t=1}^{k} \phi_t \left(\|B_t\|^2 - 1 \right) \tag{10}$$

where ϕ_t is a positive scalar indicating the Lagrange multiplier. Based on the derivation to B, we can obtain the closed form solution as:

$$B = XS^T \left(SS^T + \varphi \right)^{-1} \tag{11}$$

where φ is a diagonal matrix with its t-th entity being $\varphi_{tt} = \phi_t$. And it can be optimized through the Lagrange dual problem $\min_{\varphi_{tt} \geq 0} Tr\left(XS^T (SS^T + \varphi)^{-1} SX^T \right) + Tr(\varphi)$, which is easily solved using conjugate gradient. The whole procedure of the proposed image clustering method is clearly summarized in Algorithm 1.

Algorithm 1 Partially **T**agged **I**mage **C**lustering (**PTIC**)

Input:

Visual feature X, partially observed tag matrix T, the latent dimensionality of S and the number of clusters;

1: Initialize B and C by random matrices.
2: **while** not converge **do**
3: Fix B and C, update S using (6);
4: Fix S and B, update C using (8);
5: Fix S and C, update B using (11);
6: **end while**
7: Perform the k-means clustering algorithm on S.

Output:

Image groups based on the preset number of clusters

4. EXPERIMENTS

4.1. Evaluation datasets

Pascal VOC 2007 dataset[2]: It consists of 20 categories with a total of 9,963 image-tag pairs. We use the Color feature as the visual representation and there are a total of 399 tags. Furthermore, those image-tag pairs with multiple categories are removed. Finally we have 5,649 image-tag pairs with 30 images have no tags.

NUS WIDE dataset[3]: The database is collected from Flickr and it consists of 269,648 images in 81 categories. Six types of low level features are extracted and the 500D bag of words description is utilized as the visual feature here. As for the tag, the 1,000D tag collection is used. We select the first ten categories with each class consisting of 200 images as a subset to evaluate the proposed method. Note that in this database, 723 images have no tag information.

MIR Flickr dataset[4]: It has 15,000 image-tag pairs distributed in 38 categories. The authors provide seven types of low-level features and 2,000D most frequently used tags. The 960D GIST feature is employed as the visual feature here. Furthermore, we select 5 categories with the largest numbers of images as a subset for the experiments. In total, we have 7,933 image-tag pairs with 1,355 images have no tags.

4.2. Experimental settings

We compare our model with the baselines "k-means" and "Sparse Coding" that use no tag information, and representative works "PairwiseSC", "CentroidSC" [15] and "PVC" [18] utilizing both visual and tag features. For the methods 'PairwiseSC" and "CentroidSC", we follow [15] and choose the mean value of the Euclidean distance between all data points as the standard deviation for constructing the Gaussian kernel. As for "PVC", which is designed for incomplete feature representations, is implemented using the code released by authors.

For our method "PTIC", the dimension of sparse coding is chosen to be 300 for the three datasets and we will test its influence in the parameter selection part. Besides, we empirically assign the values of the importance matrix to be 1 and 0.01 in all the experiments like in [20]. As k-means is used in all the experiments, it is run 20 times with random initialization. Two widely used metrics, i.e., the accuracy (ACC) and the normalized mutual information (NMI), are utilized to measure the clustering performance. Readers may refer to [21] for more details about their definitions.

4.3. Experimental results

Table 1 shows the clustering accuracy and normalized mutual information of different methods on the three databases. Overall, it can be seen that our method outperforms all the compared methods. Since tag information has much higher semantic representation than that of visual features, "k-means" and "Sparse Coding" algorithms obtain much worse results than the other methods using tag information.

"PairwiseSC" and "CentroidSC" aim to find a latent space that makes the visual and tag representations being similar, and this will harm the learning process if some of the images have no tag at all. In constrast, our model utilizes tags to assist the learning process of the sparse representations, which may be less affected when confronting the scenario that the tag information is incomplete.

As for "PVC", it learns a unified latent representation for data points having complete visual and tag features based on

[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/

[3] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

[4] http://www.cs.toronto.edu/ nitish/multimodal/index.html

Methods	VOC		NUS		MIR	
	ACC (%)	NMI (%)	ACC (%)	NMI (%)	ACC (%)	NMI (%)
k-means	12.13 (0.27)	6.11 (0.11)	19.95 (0.85)	6.51 (0.29)	31.66 (0.03)	5.69 (0.01)
Sparse Coding[19]	15.22 (0.71)	6.08 (0.42)	20.13 (0.82)	6.58 (0.36)	32.31 (0.65)	6.04 (0.16)
PairwiseSC[15]	53.20 (1.63)	52.23 (1.31)	38.62 (0.67)	26.00 (0.63)	41.17 (0.59)	9.26 (0.15)
CentroidSC[15]	50.76 (2.47)	49.86 (2.08)	38.51 (1.84)	31.64 (1.16)	41.49 (0.03)	8.48 (0.01)
PVC [18]	52.97 (2.20)	51.51 (1.71)	31.08 (1.64)	23.05 (1.16)	35.71 (1.47)	6.65 (0.74)
PTIC	**56.56 (2.38)**	**53.37 (0.93)**	**42.06 (1.81)**	**34.57 (1.13)**	**43.13 (0.04)**	**9.89 (0.02)**

Table 1. Clustering results on the VOC, NUS and MIR databases. Numbers in parentheses are the std. deviations.

non-negative matrix factorization, which is effective for dealing with the text data. Compared with "PVC", we can obtain the salient structures through sparse coding on the visual feature and assist the learning process through matrix factorization on the tag feature, which is more suitable for partially tagged image clustering.

4.4. Results of partially observed tag information

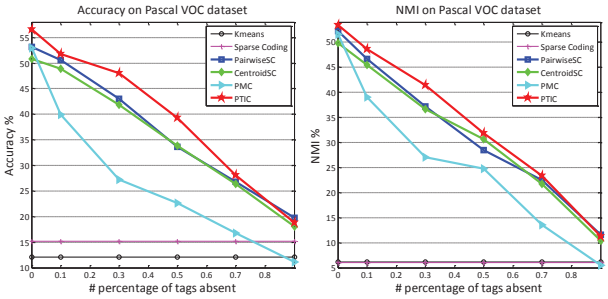

Fig. 1. Clustering results when some images have partially observed tags on the VOC dataset.

In this section, we evaluate the performance of our method facing the situation that some images have partially observed tags, which is different from Section 4.3 in which some images have no tags at all. This is also a practical scenario because the users may omit some tags when annotating images. To mimic such a scenario, we randomly remove a certain percentage of tags. We only report the results on the VOC database due to space limitation here and the other two datasets show similar results. From Figure 1, it can be seen that our model performs better with the increasing percentages of tags removed. This may be because the importance matrix we use can alleviate the tag missing situation to some degrees. As most of the tags being further removed, our model has no prominent improvements over the other methods, and this is reasonable since the tag information is too scarce to be good complementary information.

4.5. Parameter selection

In our model, λ balances the sparse coding of visual features and the matrix factorization for partially observed tag features. It is empirically selected through searching at the

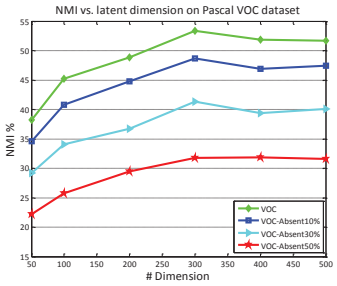

Fig. 2. NMI vs. dimension on the VOC dataset.

interval [1, 20]. As for the regularizer parameter α, it is chosen following the rules of the SLEP package. Here, we test the clustering performance vs. the latent dimension of sparse representations. To save space, only the results on the VOC dataset are reported, and other two databases show similar results. In Figure 2, VOC-Absent10%, VOC-Absent30% and VOC-Absent50% mean 10%, 30% and 50% percentages of tags are removed on the VOC dataset respectively. As the dimension of sparse representations increases, more information can be embedded and thus better clustering performance can be obtained. However, when the dimension is large enough, the clustering results keep steady because of the saturated representation ability of features.

5. CONCLUSION

In this paper, we have proposed a novel image clustering method that utilizes partially observed tags as complementary information. By enforcing both the sparse representation and the learned latent tag representation to be consistent with the partial image-tag observation, we can learn better image representations for final clustering. To this end, we have also developed an effective iterative optimization algorithm. Extensive experiments have demonstrated the effectiveness of our proposed method.

6. ACKNOWLEDGMENTS

This work is jointly supported by National Basic Research Program of China (2012CB316300), and National Natural Science Foundation of China (61175003, 61202328, 61420106015, U1435221, 61403390).

7. REFERENCES

[1] Alex Rodriguez and Alessandro Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, pp. 1492–1496, 2014.

[2] Jile Zhou, Guiguang Ding, and Yuchen Guo, "Latent semantic sparse hashing for cross-modal similarity search," *In International Conference on Research and Development in Information Retrieval*, pp. 415–424, 2014.

[3] Hao Ma, Jianke Zhu, Michael R. Lyu, and Irwin King, "Bridging the semantic gap between image contents and tags," *Transactions on Multimedia*, vol. 12, pp. 462–473, 2010.

[4] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen, "Hierarchical clustering of www image search results using visual, textual and link information," *In International World Wide Web Conferences*, pp. 952–959, 2004.

[5] Symeon Papadopoulos, Christos Zigkolis, Giorgos Tolias, Yannis Kalantidis, Phivos Mylonas, Yiannis Kompatsiaris, and Athena Vakali, "Image clustering through community detection on hybrid image similarity graphs," *In International Conference on Image Processing*, pp. 2353–2356, 2010.

[6] Symeon Papadopoulos, Christos Zigkolis, Giorgos Tolias, Yannis Kalantidis, Phivos Mylonas, Yiannis Kompatsiaris, and Athena Vakali, "Web image co-clustering based on tag and image content fusion," *In International Conference on Network Infrastructure and Digital Content*, pp. 378–382, 2010.

[7] Pierre-Aiain Moellic, Jean-Emmanuel Hauqeard, and Guillaume Pitel, "Image clustering based on a shared nearest neighbors approach for tagged collections," *In International Conference on Image and Video Retrieval*, pp. 269–278, 2008.

[8] Manjeet Rege, Ming Dong, and Jing Hua, "Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering," *In International World Wide Web Conferences*, pp. 317–326, 2008.

[9] Jinye Peng, Yi Shen, and Jianping Fan, "Cross-modal social image clustering and tag cleansing," *Journal of Visual Communication and Image Representation*, vol. 24, pp. 895–910, 2013.

[10] Shiliang Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, pp. 2031–2038, 2013.

[11] Yuhong Guo, "Convex subspace representation learning from multi-view data," *In AAAI Conference on Artificial Intelligence*, pp. 387–393, 2013.

[12] Xiaowen Dong, Pascal Frossard, Pierre Vandergheynst, and Nikolai Nefedov, "Clustering on multi-layer graphs via subspace analysis on grassmann manifolds," *Transactions on Image Processing*, vol. 62, pp. 905–918, 2014.

[13] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," *In AAAI Conference on Artificial Intelligence*, pp. 2149–2155, 2014.

[14] Mingjie Qian and Chengxiang Zhai, "Unsupervised feature selection for multi-view clustering on text-image web news data," *In International Conference on Information and Knowledge Management*, pp. 1963–1966, 2014.

[15] Abhishek Kumar, Piyush Rai, and Hal Daum Iii, "Co-regularized multi-view spectral clustering," *In International Conference on Machine Learning*, pp. 1413–1421, 2011.

[16] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han, "Multi-view clustering via joint nonnegative matrix factorization," *In SIAM International Conference on Data Mining*, pp. 252–260, 2013.

[17] Anusua Trivedi, Hal Daum III, and Scott L. DuVall, "Multiview clustering with incomplete views," *In NIPS Workshop on Machine Learning for Social Computing*, 2010.

[18] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou, "Partial multi-view clustering," *In AAAI Conference on Artificial Intelligence*, pp. 1968–1974, 2014.

[19] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng, "Efficient sparse coding algorithms," *In Advances in Neural Information Processing Systems*, pp. 801–808, 2007.

[20] Qifan Wang, Luo Si, and Dan Zhang, "Learning to hash with partial tags: Exploring correlation between tags and hashing bits for large scale image retrieval," *Transactions on Multimedia*, pp. 378–392, 2014.

[21] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and Edward Y. Chang, "Parallel spectral clustering in distributed systems," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 568–586, 2010.