Visual words assignment on a graph via minimal mutual information loss

Yue Deng
http://media.au.tsinghua.edu.cn/dengyue.html
Yanjun Qian

Yipeng Li

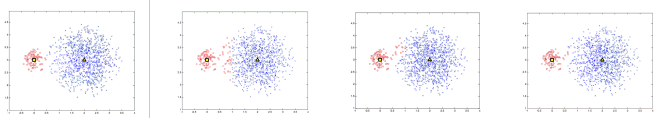
Qionghai Dai
http://media.au.tsinghua.edu.cn/qhdai.html
Guihua Er

Department of Automation
Tsinghua University
Beijing, China

Visual codewords assignment plays an important role in many Bag of Features (BoF) models for image understanding and visual recognition. It allocates image descriptors to the most similar codewords in the pre-configured visual dictionary to generate descriptive histogram for the consequent categorization. Nevertheless, existing assignment approaches, *e.g.* nearest neighbors strategy and Gaussian similarity, suffer from two problems:1) too strong Euclidean assumption and 2) neglecting the label information of the local features. Accordingly, in this paper, we propose an assignment method to simultaneously consider the above two issues in a unified model via graph learning and information theoretic criterions.

Our contributions are two-folds: 1) We propose a new local feature assignment method from the new perspective of graph learning that enables the usage of Non-Euclidean graph metric, e.g. geodesic distance and commute time for feature assignment and 2) we introduce the information theoretic penalty to reveal both the relationship of local features and their category labels. Our model exhibits both the advantages of graph learning and information theoretic learning and thus it is named Graph Assignment with minimal Mutual Information Loss (GAMIL). First of all, we describe how to assign image features via a graph.

We define the local image features set as $S = \{(f_1,l_1),(f_2,l_2)...(f_n,l_n)\}$, where $f_i \in \mathbb{R}^p$ is the image feature, *e.g.* dense sift, extracted on the original image and $l_i \in \{1,2,...C\}$ is the category label of the image that f_i is extracted from. C is the number of image categories. Therefore, in this paper, we propose to use a graph to model the samples in S. Using the manifold structure, it is possible to model the linearity among data by the locality similarity; and the global nonlinearity can be evaluated by some graph metric on the manifold, e.g. geodesic distance and commute time [3]. But the above-motioned method on a graph is only suitable to the in-sample features. For practical usage of codeword assignment, it is desirable to extend the assignment ability to the out-of-sample data. Inspired by [1], we propose to embed the graph into an Euclidean space with a linear projection matrix. In the embedded space, the original graph metric is well preserved by the Euclidean distance. Besides, it worths noting that each feature in the set S also contains the label information. During training, we know where the image feature comes from. Accordingly, the problem changes to be how to evaluate the relationship between the features F and their labels L. Fortunately, owing to the previous work [2], we know that the relationship of feature and label is always judged by the mutual information, i.e. $I(F;L)$. In probability theory and information theory, the mutual information [2] of two random variables is a quantity that measures the mutual dependence of the two random variables. It measures how much knowing one of these variables reduces uncertainty about the other. Informally, in our case, $I(F;L)$ can be interpreted as how much the uncertainty is reduced about the label L if we know the feature F. Therefore, for discriminative learning, a large mutual information score is desired. The proposed graph assignment method projects the high dimensional feature in a low dimensional space ($q < p$). Ideally we hope that the mutual information on the original graph should be kept the same in the embedding space, i.e. $I(F;L) = I(\Omega^T F,L)$. Unfortunately, reducing the dimensionality of data from high to low of course causes information loss. Therefore, instead of mutual information preservation, we propose to use the minimal mutual information loss criterion, i.e. to minimize $I(F;L) - I(\Omega^T F;L)$. Therefore, by considering both the information loss and graph similarity, the optimization for our GAMIL model is given,



(a) ground truth (b) assignment via ED. (c) assignment via GD. (d) assignment via CT.

Figure 1: A toy assignment via different methods, i.e. Euclidean Distance(ED), Geodesic Distance(GD) and Commute Time(CT).

$$\min_{\Omega} \underbrace{tr(\Omega^T F(D-W)F^T\Omega)}_{\text{graph assignment}} + \underbrace{\alpha H(L|Y=\Omega^T F)}_{\text{mutual information loss}} \quad s.t. \Omega^T FDF^T\Omega = I, \tag{1}$$

where α is a user specified parameter which trades off the graph assignment and mutual information loss. where $F = [f_1,...,f_n] \in \mathbb{R}^{p\times n}$ is the feature matrix; $\Omega \in \mathbb{R}^{p\times q}, q < p$ is the linear projection matrix. $W = [w_{ij}]$ is the weight matrix obtained on the graph which records the similarity between any two nodes i, j on the graph and $D = diag(\sum_i W_{ij})$. For an out-of-sample feature, we first project it to the subspace and then assign it to each codeword via Euclidean similarity. It is because the Euclidean distance in the embedding space represents the original nonlinear graph similarity on the manifold. For learning, the proposed model can be efficiently solved in a closed-form with the reasonable graph topology invariant approximation.

In the experiment, we randomly pick a number of images per class for training, and the left are for testing. In order to get reliable results, each experiment is repeated for 10 times (otherwise notice).

Table 1: The comparisons of GAMIL model with other state-of-the-arts on two benchmarks

Algorithms	Scene-15	Caltech-101
Hard	76.3	56.4
Soft	78.2	59.5
GAMIL	**80.7**	64.3
Info-loss[2]	74.7	-
Sparse coding[4]	80.3	**67.0**

To describe an image, we use a grid-based method to extract the dense sift features and the codebook is generated in the embedding space by K-means algorithm. For classification, we use the SVM with a histogram intersection kernel. We evaluate the proposed algorithm on two benchmarks, i.e. scene-15 and caltech-101 and our own dataset on Multiview Human Bodies (MHB). Experimental results show that our algorithm achieves state-of-the-art performances on benchmarks.

[1] Yue Deng, Qionghai Dai, Ruiping Wang, and Zengke Zhang. Commute time guided transformation for feature extraction. In *CVIU*, 2012.

[2] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *TPAMI*, 2009.

[3] Huaijun Qiu and E.R. Hancock. Clustering and embedding using commute times. In *TPAMI*, 2007.

[4] Jianchao Yang, Kai Yu, Yihong Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.