Learning Contextual Metrics for Automatic Image Annotation

Zuotao Liu[1], Xiangdong Zhou[1], Yu Xiang[1], and Yan-Tao Zheng[2]

[1] Fudan University, Shanghai, China
{082024020,xdzhou,072021109}@fudan.edu.cn
[2] Institute for Infocomm Research, Singapore
yantaozheng@gmail.com

Abstract. The semantic contextual information is shown to be an important resource for improving the scene and image recognition, but is seldom explored in the literature of previous distance metric learning (DML) for images. In this work, we present a novel Contextual Metric Learning (CML) method for learning a set of contextual distance metrics for real world multi-label images. The relationships between classes are formulated as contextual constraints for the optimization framework to leverage the learning performance. In the experiment, we apply the proposed method for automatic image annotation task. The experimental results show that our approach outperforms the start-of-the-art DML algorithms.

1 Introduction

The fundamental issue of multimedia retrieval and visual recognition is to capture the similarity between visual objects. Finding proper distance metric for similarity measure is critical for these tasks. However, distance metric is task-oriented and manually selecting distance metric is tedious and even unrealistic for various practical applications. Therefore, distance metric learning (DML), which learns distance metric by exploring the available intrinsic information from training data, draws increasing attentions from research and industry communities. In the last few years, many metric learning algorithms are proposed, such as RCA[1], NCA[2], LMNN[3], DCA[4] and ITML[5], which are shown to perform well in some classification and clustering problems. The principal of most previous work investigate side information from training data, e.g. data points are considered to be either "similar" or "dissimilar", and utilize it as pairwise constraints to learn a holistic Mahalanobis distance or linear transformation.

For the scene and image classification and semantic annotation problems, the task is to assign multiple labels to each vision instance, so called multi-label classification. It is an extension of the common single-label problem. Wu et al. [6] presented a probabilistic distance metric learning framework that can derive constraints from the uncertain side information of multi-labeled data and learn a distance metric from the derived constraints. Qi et al. [7] proposed to learn a metric that can keep the linear transformation between the visual space and

G. Qiu et al. (Eds.): PCM 2010, Part I, LNCS 6297, pp. 124–135, 2010.

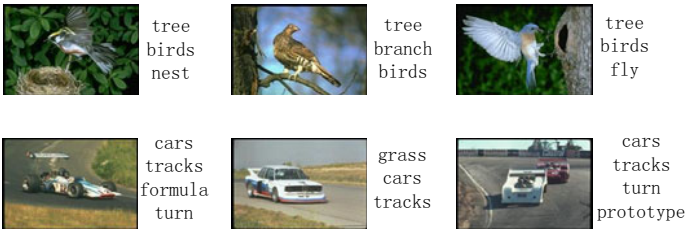

Fig. 1. Illustration of semantic context by example images from Corel image data set and their human annotations

label space, and formulates it as a semi-definite programming (SDP) problem. However, most previous work deal with the problem of holistic distance metric learning, i.e. one metric for all classes. When the number of classes of the input space grows, the holistic manner is difficult to seize the underlying characteristics of each class [8]. Specifically, in the multi-label setting, the interaction between classes and images need to be explored for metric leaning [7].

In this paper, we present a novel distance metric learning approach that can learn a set of metrics simultaneously one for each class. In the proposed method, the discriminations of different classes and the interactions between images and classes are explored integratedly. Moreover, the semantic contextual information between classes are utilized to reduce the over-fitting of the proposed method. The semantic context comes from the co-occurrence of some class labels in the same scene (image), for instance, "bird" and "tree", "car" and "track", and so on. Figure 1 presents some illustrative images of Corel data set [9] showing examples of frequently co-occurring class labels.

Semantic context is shown to be an important resource for improving the performance of vision recognition [10], but is seldom explored in the metric learning literature for images. Intuitively, for metric learning of scene and images, knowing some classes co-occurring frequently provides hint that their corresponding metrics are similar in some extent. In particular, we describe the contextual constrains by using KL-divergence between classes, which is based on the principle of bijection between the Mahalohobis distance and an equal-mean multivariate Gaussian distribution [5]. The main contributions of the work are as follows:

We propose the Contextual Metric Learning framework for multi-label images and scenes. We introduce the semantic contextual constraints into the proposed metric learning framework and apply the learning method into the automatic image annotation (AIA) task.

Our proposed learning framework can be efficiently solved with a closed-form solution to obtain a set of optimal metrics simultaneously one for each class to uncover the intrinsics of each scene class.

To demonstrate the learning ability of the proposed method, we apply our algorithm for the application of automatic image annotation on two real world data sets, Corel and a subset of TRECVID-2005 data set. The experimental results show that our algorithm outperforms the-state-of-the art algorithms.

The rest of the paper is organized as follows. In section 2, we give an brief overview of the related work; Section 3 is for our learning models. In Section 4, we give out the experiments about utilizing the learned metrics for AIA task. Section 5 concludes our work.

2 Related Work

Most of the previous work in supervised metric learning relies on learning a holistic Mahalanobis distance. ITML[5] models the DML problem in an information-theoretic setting by leveraging the relationship between the multivariate Gaussian distribution and the set of Mahalanobis distances. It formulates the problem as that of minimizing the differential relative entropy between two multivariate Gaussians under constraints on the distance function, and expresses it as a particular Bregman optimization problem-that of minimizing the LogDet divergence subject to linear constraints.

There are several algorithms that attempt to learn multiple metrics for a learning task. [8] attempted to learn different Mahalanobis distance metrics in different parts of the input space in the setting of single-labeled data, and [11] presented an algorithm that learns a few similarity category specific metrics while simultaneously grouping categories together and assigning one of these metrics to each group, however, both of these two algorithms are designed for single-label tasks. [12] attempted to learn a amount of metrics for each instance using metric propagation, which is not practical for large numbers of instances.

A significant amount of work have been devoted to the problem of AIA. Generative models[10] focus on learning the correlations between images and semantic concepts, while discriminative models[13] formulate AIA as a classification problem and apply classification techniques to AIA, such as Support Vector Machine (SVM) and Gaussian mixture model. Moreover, Zhou et al. [14] proposed a hybrid approach combining user-provided tags and image visual contents under a unified probabilistic framework. Guillaumin et al. [15] proposed a discriminative metric learning algorithm (Tagprop) for AIA, which uses Bernoulli models for each keyword and weighted nearest neighbor approach for tag prediction. The distance metric used in Tagprop is a linear combination of a set of base distances, which are derived from manually assigned metrics for different features, and the weights for them can be learnt from the training set.

3 The Contextual Metric Learning Method

In this section, we present our contextual metric learning method. We first present the problem formulation of learning multiple metrics in multi-label settings. Then we describe our regularization framework of these metrics using semantic contextual information. Finally, we detail the algorithm of training multiple metrics under the contextual constraints.

3.1 Learning Multiple Metrics

Metric learning aims to seek a Positive Semi-Definite (PSD) matrix M which parameterizes the Mahalanobis distance. To find the matrix M, some constraints must be imposed into the learning procedure. In classification, a common constraint is that instances from the same class are closer to each other than instances from different classes. However, the constraint is insufficient for multi-label classification setting, where each data point can belong to multiple classes simultaneously. For example, if two instances both belong to class A while one belongs to class B and the other does not, then the two constraints induced by the two classes will conflict. To solve the problem in multi-label metric learning, we propose a novel method to learn multiple metrics simultaneously, one for each class, while the relationships between metrics are exploited in our method.

Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ denote a set of n data points, where $\mathbf{x}_j \in \mathbb{R}^d, j = 1, \ldots, n$ is a d dimensional feature vector. We denote the label of data point $\mathbf{x}_j$ by $\mathbf{y}_j = \{y_j^1, y_j^2, \ldots, y_j^m\}$, where $y_j^i \in \{0, 1\}$ indicates whether $\mathbf{x}_j$ belongs to class $\mathcal{C}_i$ or not and m denotes the number of classes. Our goal is to learn a set of Mahalanobis matrices $M_i, i = 1, \ldots, m$ for different classes

$$d_{M_i}(\mathbf{x}_j, \mathbf{x}_k) = \sqrt{(\mathbf{x}_j - \mathbf{x}_k)^T M_i (\mathbf{x}_j - \mathbf{x}_k)} \tag{1}$$

under a set of pairwise constraints among the data points

$$\mathcal{S}_i = \{(\mathbf{x}_j, \mathbf{x}_k) | \mathbf{x}_j \in \mathcal{C}_i \text{ and } \mathbf{x}_k \in \mathcal{C}_i\} \tag{2}$$

$$\mathcal{D}_i = \{(\mathbf{x}_j, \mathbf{x}_k) | \mathbf{x}_j \in \mathcal{C}_i \text{ and } \mathbf{x}_k \notin \mathcal{C}_i\}, \tag{3}$$

where $\mathcal{S}_i$ is the set of similar pairwise constraints derived from class $\mathcal{C}_i$ and $\mathcal{D}_i$ is the corresponding set of dissimilar pairwise constraints. We define a loss function for each Mahalanobis matrix, which minimizes the distances in the similar constraints and maximizes the distances in the dissimilar constraints:

$$\begin{aligned} L(M_i, \mathcal{S}_i, \mathcal{D}_i) = & \frac{\gamma_s^i}{2} \sum_{(\mathbf{x}_j, \mathbf{x}_k) \in \mathcal{S}_i} (\mathbf{x}_j - \mathbf{x}_k)^T M_i (\mathbf{x}_j - \mathbf{x}_k) \\ & - \frac{\gamma_d^i}{2} \sum_{(\mathbf{x}_j, \mathbf{x}_k) \in \mathcal{D}_i} (\mathbf{x}_j - \mathbf{x}_k)^T M_i (\mathbf{x}_j - \mathbf{x}_k) \end{aligned} \tag{4}$$

where γ_s^i and γ_d^i are two parameters which can balance the tradeoff between similar and dissimilar constraints. Finally, we combine the loss functions for all the Mahalanobis matrices to obtain the loss function of our contextual metric learning framework:

$$L(\mathbf{M}, \mathbf{S}, \mathbf{D}) = \sum_{i=1}^{m} L(M_i, \mathcal{S}_i, \mathcal{D}_i), \tag{5}$$

where $\mathbf{M} = \{M_i | i = 1, 2, \ldots, m\}$ denotes the matrices to be learnt, $\mathbf{S} = \{S_i | i = 1, 2, \ldots, m\}$ denotes the set of similar pairwise constraints and $\mathbf{D} = \{D_i | i = 1, 2, \ldots, m\}$ denotes the set of dissimilar pairwise constraints.

3.2 Contextual Regularization

We add a regularization term into our metric learning framework to incorporate our prior knowledge about the task and prevent the learnt matrices from over-fitting. In many settings, we require the learnt metric to be close to some given Mahalanobis distance function. For example, if the data is Gaussian, we regularize the Mahalanobis matrix by the inverse of the sample covariance. In some settings, Euclidean distance may work well. In our method, we regularize the matrices by both the Euclidean distance and the relationships between themselves.

As noted by ITML [5], there exists a bijection between a Mahalanobis distance and an equal-mean multivariate Gaussian distribution. Given a Mahalanobis matrix M, the corresponding multivariate Gaussian distribution can be expressed as $P(\mathbf{x}; \boldsymbol{\mu}, M) = \frac{1}{Z} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T M(\mathbf{x} - \boldsymbol{\mu}))$, where Z is a normalizing constant and M equals to the inverse of the covariance of the distribution. So we can measure the distance between two Mahalanobis distance functions by the differential relative entropy between their corresponding multivariate Gaussians:

$$KL\big(P(\mathbf{x}; \boldsymbol{\mu}_i', M_{i'}) \| P(\mathbf{x}; \boldsymbol{\mu}_i, M_i)\big) = \int P(\mathbf{x}; \boldsymbol{\mu}_i', M_{i'}) \log \frac{P(\mathbf{x}; \boldsymbol{\mu}_i', M_{i'})}{P(\mathbf{x}; \boldsymbol{\mu}_i, M_i)} d\mathbf{x}. \quad (6)$$

In our metric learning method, we regularize each metric by Euclidean distance function. So we add the following single regularization term into our framework:

$$\sum_{i=1}^{m} KL\big(P(\mathbf{x}; \boldsymbol{\mu}_i, M_0) \| P(\mathbf{x}; \boldsymbol{\mu}_i, M_i)\big), \quad (7)$$

where $M_0 = I$ is the identity matrix. For multi-label distance metric learning, all the multi-labeled instances are presented as mixed features of different concepts. So if two classes co-occur frequently, they share the similar distribution, and their KL-divergence tends to be small. So we regularize the matrices by the contextual relationships between classes. We add the following pairwise regularization term into our framework:

$$\sum_{i=1}^{m} \sum_{i' \in \mathcal{N}_i} KL\big(P(\mathbf{x}; \boldsymbol{\mu}_i', M_{i'}) \| P(\mathbf{x}; \boldsymbol{\mu}_i, M_i)\big), \quad (8)$$

where $\mathcal{N}_i$ is the set of classes correlated with class i. We measure the correlations between classes based on the co-occurrences of classes in a training data set. Two classes co-occur if they are associated with the same instance in the training set. We define a correlation measure between classes by:

$$P(\mathcal{C}_{i'} | \mathcal{C}_i) = \frac{|\mathcal{C}_i \bigcap \mathcal{C}_{i'}|}{|\mathcal{C}_i|}, \quad (9)$$

which is the estimation of the prior conditional probability of observing class $\mathcal{C}_{i'}$ on condition of class $\mathcal{C}_i$. Based on the above measure, we construct a graph

structure for classes and define class i' is a neighbor of class i, i.e. $i' \in \mathcal{N}_i$, if and only if $P(\mathcal{C}_{i'}|\mathcal{C}_i) \geq P_0$, $\forall i, i' = 1, 2, \ldots, m$, where P_0 is a predefined threshold constant and the value is set to 0.1 in the experiment. The constructed neighborhood system is not symmetric since the interaction between two classes is not mutually equal.

By combining Eq. (7) and Eq. (8), we obtain the regularization term of our method:

$$R(\mathbf{M}) = \sum_{i=1}^{m} KL\big(P(\mathbf{x}; \boldsymbol{\mu}_i, M_0) \| P(\mathbf{x}; \boldsymbol{\mu}_i, M_i)\big) + \lambda \sum_{i=1}^{m} \sum_{i' \in \mathcal{N}_i} KL\big(P(\mathbf{x}; \boldsymbol{\mu}'_i, M_{i'}) \| P(\mathbf{x}; \boldsymbol{\mu}_i, M_i)\big), \tag{10}$$

where λ is a constant controlling the tradeoff between the single regularization and the pairwise regularization. The above regularization utilizes the contextual relationships between classes. So we refer to it as *Contextual Regularization*.

3.3 Algorithm

By combining the loss function (5) and the regularization (10) , we obtain our objective function of CML framework:

$$L'(\mathbf{M}, \mathbf{S}, \mathbf{D}) = L(\mathbf{M}, \mathbf{S}, \mathbf{D}) + R(\mathbf{M}). \tag{11}$$

We learn the metrics by minimizing the above loss function under PSD constraints, which is equivalent to the following optimization problem

$$\begin{aligned} &\min_{M_i \succeq 0, i=1,\ldots,m} L'(\mathbf{M}, \mathbf{S}, \mathbf{D}) \\ &= \min_{M_i \succeq 0, i=1,\ldots,m} \sum_{i=1}^{m} L(M_i, \mathcal{S}_i, \mathcal{D}_i) + \sum_{i=1}^{m} KL\big(P(\mathbf{x}; \boldsymbol{\mu}_i, M_0) \| P(\mathbf{x}; \boldsymbol{\mu}_i, M_i)\big) \\ &+ \lambda \sum_{i=1}^{m} \sum_{i' \in \mathcal{N}_i} KL\big(P(\mathbf{x}; \boldsymbol{\mu}'_i, M_{i'}) \| P(\mathbf{x}; \boldsymbol{\mu}_i, M_i)\big) \end{aligned} \tag{12}$$

In the loss function (5), if we define

$$K_i^{jk} = \begin{cases} \gamma_s^i, \text{ if } (\mathbf{x}_j, \mathbf{x}_k) \in \mathcal{S}_i \\ -\gamma_d^i, \text{ if } (\mathbf{x}_j, \mathbf{x}_k) \in \mathcal{D}_i, \end{cases} \tag{13}$$

then we can get

$$\begin{aligned} L(M_i, \mathcal{S}_i, \mathcal{D}_i) &= \frac{1}{2}\sum_{j,k=1}^{n}(\mathbf{x}_j - \mathbf{x}_k)^T M_i(\mathbf{x}_j - \mathbf{x}_k)K_i^{jk} \\ &= \sum_{j,k=1}^{n}(\mathbf{x}_j^T M_i \mathbf{x}_j - \mathbf{x}_j^T M_i \mathbf{x}_k)K_i^{jk} \\ &= tr(XM_iX^TD_i) - tr(XM_iX^TK_i) \\ &= tr(XL_iX^TM_i) \end{aligned} \tag{14}$$

where D_i is a diagonal matrix whose diagonal are the sums of the row elements of K_i, and $L_i = D_i - K_i$ is the Laplacian matrix of K_i. $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ is a matrix composed of all the training instances. It has been shown in [16] that the differential relative entropy between two multivariate Gaussians can be expressed as the convex combination of a Mahalanobis distance between mean vectors and the LogDet divergence between the covariance matrices:

$$\begin{aligned} &KL\big(P(\mathbf{x}; \boldsymbol{\mu}_i', M_{i'}) \| P(\mathbf{x}; \boldsymbol{\mu}_i, M_i)\big) \\ &= \frac{1}{2}D_{ld}(M_i, M_{i'}) + \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})^T M_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}), \end{aligned} \tag{15}$$

where LogDet divergence $D_{ld}(M_i, M_{i'})$ is a Bregman matrix divergence generated by the convex function $\phi(X) = -\log\det X$ defined over the cone of positive-definite matrices, and it equals to (for $d \times d$ matrices M_i and $M_{i'}$)

$$D_{ld}(M_i, M_{i'}) = tr(M_i M_{i'}^{-1}) - \log\det(M_i M_{i'}^{-1}) - d. \tag{16}$$

By substituting Eq. (14) and Eq. (15) into Eq. (12), we can get the following optimization problem for our CML framework:

$$\begin{aligned} &\min_{M_i \succeq 0, i=1,\ldots,m} L'(\mathbf{M}, \mathbf{S}, \mathbf{D}) \\ &= \min_{M_i \succeq 0, i=1,\ldots,m} \sum_{i=1}^{m} D_{ld}(M_i, I) + \lambda \sum_{i=1}^{m}\sum_{i' \in \mathcal{N}_i} \big(\frac{1}{2}D_{ld}(M_i, M_{i'}) \\ &+ \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})^T M_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})\big) + \sum_{i=1}^{m} tr(XL_iX^TM_i). \end{aligned} \tag{17}$$

We solve the above optimization problem (17) by alternating optimization strategy, where we iteratively optimize each matrix on condition of fixing the other matrices. When fixing all the matrices $M_{i''}$, $i'' \neq i$, the optimization for M_i is a standard formulation of Semi-Definite Programming (SDP) [17], which can be solved using existing convex optimization packages. We iteratively optimize all the M_i until convergence.

Considering the expensive time cost of SDP problem, we adopt a closed-form approximative solution that can be obtained efficiently by taking the derivative of Eq. (17):

$$\frac{\partial L'(\mathbf{M},\mathbf{S},\mathbf{D})}{\partial M_i} = I - M_i^{-1} + \frac{\lambda}{2}\sum_{i'\in\mathcal{N}_i}(M_{i'}^{-1} - M_i^{-1} + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})^T) + XL_iX^T, \forall i = 1, 2, \ldots, m. \quad (18)$$

By setting the above derivative to zero, we can get

$$(1+\frac{\lambda}{2}n_i)M_i^{-1} - \frac{\lambda}{2}\sum_{i'\in\mathcal{N}_i}M_{i'}^{-1} = \frac{\lambda}{2}\sum_{i'\in\mathcal{N}_i}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})^T + XL_iX^T + I, \quad (19)$$

where n_i is the number of neighbors of class i, i.e. $|\mathcal{N}_i|$. If we define

$$\widetilde{Y}_i = \frac{\lambda}{2}\sum_{i'\in\mathcal{N}_i}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})^T + XL_iX^T + I \quad (20)$$

$$\widetilde{H} = (diag(1 + \frac{\lambda}{2}n_i)_{m\times m} - \frac{\lambda}{2}N)^{-1}, \quad (21)$$

where $diag(1+\frac{\lambda}{2}n_i)_{m\times m}$ is a diagonal matrix with the ith element $1+\frac{\lambda}{2}n_i$, and N is the adjacency matrix derived from the class graph, $N_{ii'} = 1$ if and only if $i' \in \mathcal{N}_i$. Then the minimum point of objective function (17) can be derived as

$$M_i = \Big(\sum_{i'=1}^{m}\widetilde{H}_{ii'}\widetilde{Y}_{i'}\Big)^{-1} \quad (22)$$

In practice, we can control the parameter λ to ensure the learnt matrices $M_i, i = 1, 2, \ldots, m$ to be PSD.

4 Experiments

We apply metric learning algorithms for AIA task on two commonly used benchmarks: Corel and TRECVID-2005.

4.1 Experimental Dataset

Corel Dataset: The Corel image data set [9] contains 5,000 images each of which is labeled with 1-5 keywords, and there are totally 374 keywords used in the data set. Because most of the keywords only have few positive samples, we train metrics for the most popular 70 keywords which have 60 positive samples at least. For the subset of the largest 70 keywords, we get 4431 images for training and 490 images for testing. Each image is annotated by average 2.65 labels.

TRECVID-2005 Dataset: TheTRECVID-2005 data set contains about 108 hours of multi-lingual broadcast news, which is more diverse and represents the real world scenario. Compared with Corel, TRECVID data set provides more positive samples for each concept, and the concept space is smaller. Following the work of [13], we select training and testing data from 90 videos and the other 47 videos respectively. For each concept, we randomly select no more than 500 and 100 positive samples for training and testing respectively. As a result, we have 6,657 key-frames for training and 1,748 key-frames for testing.

4.2 Image Representation

We extract 5 different kinds of features commonly used for image classification and retrieval. We use two types of global features: Gist features [18] and color histograms. The color histograms are calculated with 8 bins in each color channel for RGB, LAB and HSV representations, which results in three 512-dimensional feature vectors for each image. For local features, we use SIFT and adopt the soft-weighting scheme [19] for bag-of-features of 500 dimensions. Considering the efficiency of Mahalanobis metric, we apply PCA[20] to reduce the dimension of obtained features. On Corel dataset, each kind of features are reduced to 10-dimensional vectors and we get a 50-dimensional vector for each image, while each kind of features are reduced to 20-dimensional vectors and each image is presented as a 100-dimensional vector on TRECVID-2005 dataset.

4.3 Experimental Setup and Evaluation Measures

The parameter setting of CML is as follows: λ of Eq. 10 is set to 0.1, which balances the single regularization and the contextual regularization. We set $\gamma_s^i = 1/n_p^i$ and $\gamma_d^i = 0$ in Eq. (13), where n_p^i is the number of derived similar constraints from ith class. For AIA task, we employ weighted nearest neighbor approach for tag prediction. The tag presence prediction for image j is a weighted sum over the nearest neighbor training images, indexed by k:

$$p(y_j^i = +1) = \sum_k \pi_{jk}^i p(y_k^i = +1|k), \tag{23}$$

where π_{jk}^i denotes the distance based weight of image k for predicting the tags of image j, which can exploit the effectiveness of distance metrics sufficiently:

$$\pi_{jk}^i = \frac{\exp(-w_i \, d_{M_i}(j,k))}{\sum_{k'} \exp(-w_i \, d_{M_i}(j,k'))}. \tag{24}$$

w_i of above equation is a parameter to control the decay of weights and can be learned from the training set for each class respectively by using the loss function of [15].

We use recall, precision and F1 score for fixed annotation length to evaluate the performance of AIA of the metric learning methods. For a given query word w, let $|W_G|$ be the number of human annotated images with label w in the test set, $|W_M|$ be the number of annotated images with the same label of the annotation algorithm, and $|W_C|$ be the number of correct annotations of our algorithm, then recall, precision and F1 score are defined as $recall = \frac{|W_C|}{|W_G|}$, $precision = \frac{|W_C|}{|W_M|}$ and $F1 = \frac{2 \times recall \times precision}{recall + precision}$. We compute recall and precision for each keyword and then average them to measure the annotation performance.

4.4 Experimental Results

Comparisons of AIA Performance. The Euclidean distance and the state-of-the-art metric learning algorithm ITML [5] are adopted for comparison in the

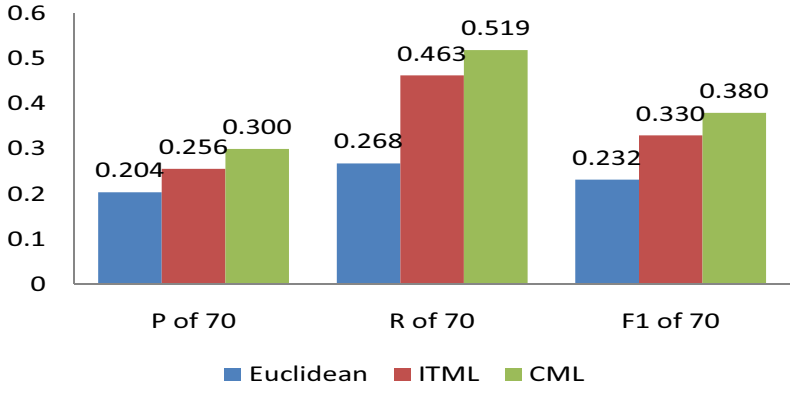

Fig. 2. Annotation performance comparison with Euclidean distance and ITML on Corel dataset. R, P, F1 denote the average precision, average recall and F1 score respectively.

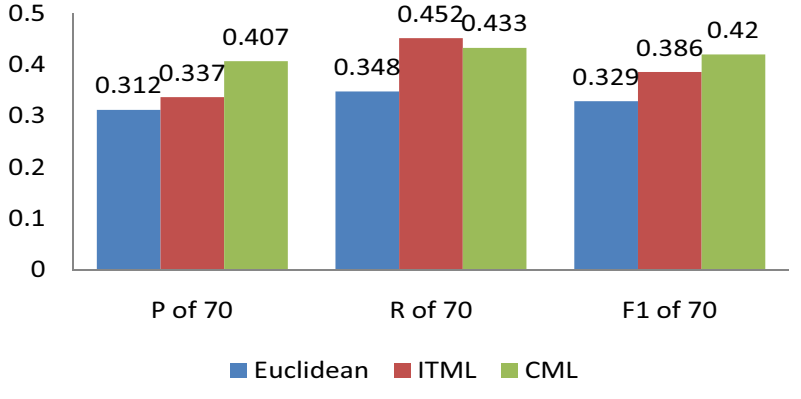

Fig. 3. Annotation performance comparison with Euclidean distance and ITML on TRECVID-2005 dataset

experiment. We learn metrics for each class using ITML and CML respectively, and apply the obtained metrics for AIA. For ITML, we select 5000 constraints for each class, where 1000 of which are similar constraints and others are dissimilar. For learning all the 70 distance metrics of Corel dataset, our closed-from solution needs only 47 seconds on Pentium 4 computer platform, whereas ITML needs more than 1 hour. According to the widely used protocol, we use the 5 most relevant keywords to annotate each Corel test image, and use the 6 most relevant keywords to label each TRECVID-2005 test image. The experimental results are shown in Figure 2 and Figure 3 respectively. From Figure 2 we can draw several observations. Firstly, metric learning can help AIA task significantly. Secondly, our model achieves the best performance on Corel dataset. Compared

with ITML, CML improves the average precision by 17.2%, the average recall by 12.1%. From Figure 3, we can see although ITML gets a better average recall, our model gets the best average precision of TRECVID-2005 and outperforms ITML by 8.8% on F1 score.

Evaluation of Semantic Context. To further evaluate the effectiveness of the semantic context for metric learning, we set different values of λ in Eq. (17) to compare the AIA performance of our CML. The annotation results of Corel and TRECVID-2005 are shown in Table 1, where $\lambda = 0$ means no context is used in CML. The table shows that with context constrains the AIA performance are improved on both data sets. It also shows that the effectiveness of context constrains in TRECVID-2005 is bigger than that in Corel. From the table, it can be observed that the semantic graph derived from Corel is more sparse, e.g. the average number of neighbors for each site in the semantic graph is only 4.97. Whereas, on our TRECVID-2005 data set, there are on average 10.84 keywords for each site. Therefore, it meas more contextual constrains taking part in the procedure of metric learning on TRECVID-2005 data set and resulting better performance, e.g. 10.2% higher in the F1 score compared with CML without contextual constraints.

Table 1. The AIA performance of CML with different semantic context settings

Dataset	Corel (70 keywords)		TRECVID-2005(39 keywords)	
Average #neighbors	4.97		10.84	
Models	CML($\lambda = 0.1$)	CML($\lambda = 0$)	CML($\lambda = 0.1$)	CML($\lambda = 0$)
Precision	**0.300**	0.292	**0.407**	0.364
Recall	**0.519**	0.515	**0.433**	0.399
F1 score	**0.380**	0.373	**0.420**	0.381

5 Conclusion

In this work, we investigate the contextual information between classes to improve the performance of distance metric learning and apply it to automatic image annotation. The intuition is that one distance metric for each class by exploiting contextual information for vision recognition is more close to the ways how people recognize visual objects. We report experimental results on two real world data sets and show that our method perform well for the task of AIA. For the future work, we will extend the contextual constrains by introducing multiple context to further leverage the power of our learning framework and apply it to scene and video classification and annotation.

Acknowledgment

This work was partially supported by the Natural Science Foundation of China under Grant No.60773077.

References

1. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning distance functions using equivalence relations. In: ICML (2003)
2. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: NIPS (2004)
3. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: NIPS (2005)
4. Hoi, S.C., Liu, W., Lyu, M.R., Ma, W.Y.: Learning distance metrics with contextual constraints for image retrieval. In: CVPR (2006)
5. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML (2007)
6. Wu, L., Hoi, S.C., Jin, R., Zhu, J., Yu, N.: Distance metric learning from uncertain side information with application to automated photo tagging. In: ACM MM (2009)
7. Qi, G.J., Hua, X.S., Zhang, H.J.: Learning semantic distance from community-tagged media collection. In: ACM MM (2009)
8. Weinberger, K.Q., Saul, L.K.: Fast solvers and efficient implementations for distance metric learning. In: ICML (2008)
9. Duygulu, P., Barnard, K., de Freitas, J., Forsyth, D.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
10. Xiang, Y., Zhou, X., Chua, T.S., Ngo, C.W.: A revisit of generative model for automatic image annotation using markov random fields. In: CVPR (2009)
11. Babenko, B., Branson, S., Belongie, S.: Similarity metrics for categorization: from monolithic to category specific. In: ICCV (2009)
12. Zhan, D.C., Li, M., Li, Y.F., Zhou, Z.H.: Learning instance specific distance using metric propagation. In: ICML (2009)
13. Xiang, Y., Zhou, X., Liu, Z., Chua, T.S., Ngo, C.W.: Semantic context modeling with maximal margin conditional random fields for automatic image annotation. In: CVPR (2010)
14. Zhou, N., Cheung, W., Xue, X.Y., Qiu, G.: Collaborative and content-based image labeling. In: ICPR (2008)
15. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV (2009)
16. Daivs, J.V., Dhillon, I.: Differential entorpic clustering of multivariate gaussians. In: NIPS (2006)
17. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press, Cambridge (2003)
18. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the patial envelop. IJCV 42, 145–175 (2001)
19. Jiang, Y., Ngo, C., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: CIVR (2007)
20. Fukunaga, K.: Introduction to statistical pattern recognition. Elsevier, Amsterdam (1990)