Unsupervised Metric Fusion by Cross Diffusion

Bo Wang[1], Jiayan Jiang[2], Wei Wang[4], Zhi-Hua Zhou[4], and Zhuowen Tu[2,3]
[1]Department of Computer Science, University of Toronto
[2]Lab of Neuro Imaging, University of California, Los Angeles
[3]Microsoft Research Asia
[4]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210046, China

Abstract

Metric learning is a fundamental problem in computer vision. Different features and algorithms may tackle a problem from different angles, and thus often provide complementary information. In this paper, we propose a fusion algorithm which outputs enhanced metrics by combining multiple given metrics (similarity measures). Unlike traditional co-training style algorithms where multi-view features or multiple data subsets are used for classification or regression, we focus on fusing multiple given metrics through diffusion process in an unsupervised way. Our algorithm has its particular advantage when the input similarity matrices are the outputs from diverse algorithms. We provide both theoretical and empirical explanations to our method. Significant improvements over the state-of-the-art results have been observed on various benchmark datasets. For example, we have achieved 100% *accuracy (no longer the bull's eye measure) on the MPEG-7 shape dataset. Our method has a wide range of applications in machine learning and computer vision.*

1. Introduction

Data samples are often given as high dimensional points, whereas they live in much lower intrinsic spaces (manifolds); utilizing the intrinsic data manifold structure therefore is an important topic in learning and vision [29, 21]; computing faithful manifold metrics (distance/similarity measures) leads to good performances in a wide range of applications in classification, segmentation, regression, image search/retrieval, and visualization [24, 16]. For high dimensional data, a direct approach of distance metric learning (e.g. Mahalanobis distances) [15, 36, 23, 7] is often used in the context of supervised learning.

One idea to derive a good distance measure is to explicitly construct a new embedding space with distance propagation which is more faithful to the manifold structure and hence induces a better distance notion. The same idea can be extended to semi-supervised cases, where a limited portion of data labels are given. For example, label propagation [40] and its variants [32] use a diffusion process to propagate labels to the unlabeled data samples along the manifold. One promising approach, diffusion map [5], defines a new metric, diffusion distances, between data samples; an input similarity matrix is then improved through a diffusion process.

The diffusion/propagation process improves the individual inputs; in practice, we are often given multiple measures by different algorithms/metrics, which are also complementary to each other. We therefore are facing an additional fusion [17] task on top of the metric learning problem.

From a different angle, co-training style algorithms allow classifiers trained on different views of the features [4, 6, 31, 25] or different subsets of the training data [9, 39] to pull out more samples from unlabeled data to help each other. PAC bounds were given for co-training on multi-view features [6] and single-view multi-classifiers [34]. Recently, co-regularization has been adopted for multi-view learning [31, 27]. However, there are few algorithms which address the problem in an unsupervised manner. In [38] an unsupervised Bayesian kernel was proposed to fuse the information induced from multiple views. This simple fusion technique suffers from being sensitive to the parameters in the algorithm and noise in the data. In this paper, we develop a dynamic process to fuse multiple metrics in an unsupervised way.

In terms of application, shape/image retrieval is an important topic in computer vision; due to being intrinsically high-dimensional and ambiguous, shape/image retrieval remains a challenging problem. Recent advances in this category attempts to apply transductive or semi-supervised learning [37] to enhance the retrieval results. Depending on the affinity relationship on a simple graph, semi-supervised learning techniques could be utilized to boost the image retrieval performance. However, it may not be sufficient to represent the full affinity relations by only one fixed graph.

In this paper, we propose a metric learning algorithm, *cross-diffusion process*, for generating enhanced similarity

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

measures by fusing multiple given metrics; it is particularly advantageous when the input similarity matrices are the outputs by diverse algorithms; in this case, we know the pairwise distances but no explicit features are given. The main contribution of this paper includes: (1) We tackle the ranking/retrieval problem by performing fusion through a dynamic process, *cross-diffusion process*, and we design an extension to deal with multiple input metrics. (2) We show the convergence of the cross-diffusion process and provide theoretical interpretations to cross diffusion. (3) On a variety of benchmark datasets, e.g. the MPEG-7 shape dataset, we observed significant improvement over the state-of-the-art results, reaching a near-perfect 100% direct retrieval accuracy (this is even a more difficult criterion than the traditional bull's eye measure). Fig. 1 shows an example of shape retrieval in MPEG-7 dataset. The first row shows the retrieval results by the shape descriptor "Shape Contexts"(SC) [3], and the second row is the result of another descriptor "Inner Distance" [13]. The proposed method combines these two shape descriptors and the retrieval outputs are shown in the last row. We can see our method can greatly improve the accuracy over the baseline. We focus on the image retrieval task here but our method can be applied to other learning tasks (we conducted other tasks for e.g. classifications and dimensionality reduction and also observed significant improvement over the state-of-the-art methods, due to the space limit, we choose not to elaborate on these applications).

1.1. Related Works

Graph-based approaches have been proposed to study the manifold structure defined by a set of data points. For example, label propagation [40] is used to propagate the label information to unlabeled points along the data manifold; Markov random walks are constructed for a more faithful representation of data affinities respecting the manifold structure [28]. More recently, attentions have been given to the work of diffusion maps [5], which is based on the notion of diffusion distances induced from a diffusion process.

There is also active research along the line of information fusion. The idea in the seminal work of co-training [4] is to bootstrap two conditionally independent classifiers by providing each other with labels for the unlabeled data. The existing literature on metric learning [36, 7] mostly focus on learning a Mahalanobis distance metric. Very recently, a co-training work for spectral clustering is proposed in [10], but we focus on metric learning here; we perform cross diffusion instead of clustering; there is no theoretical justification in [10].

The most related work to ours is the co-transduction method [2] which fuses two input similarity measures through transduction. However, our approach is more natural and general than co-transduction since we perform fusion altogether whereas ranking can only be performed one-by-one in [2]; significantly improved results over those by co-transduction have been observed as we will see in the experiments; since no global comparable similarity measures can be obtained by co-transduction, the application of co-transduction is rather limited.

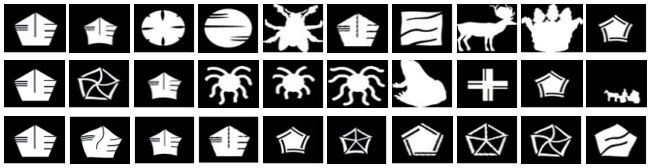


Figure 1. The first 10 retrieved shapes in MPEG7 by Shape Context (SC) [3] (first row), Inner Distance(IDSC) [13] (second row) and Cross Diffusion with SC & IDSC (last row).The first column shows the query shape.

2. Cross Diffusion

In this section, we describe our cross-diffusion method; it has a tie to the diffusion maps [5] algorithm. Due to the space limit, we refer the readers to [5] for the details.

2.1. Background

Given a finite weighted graph $G = (V, E, W)$, consisting of a set of vertices V based on the data set $X = \{x_i, i = 1, \ldots, n\}$, a subset edges E of $V \times V$, and a nonnegative symmetric weight function $W : E \to [0, 1]$. If $W(i, j) > 0$, we say that there is an edge between x_i and x_j. We interpret the weight $W(i, j)$ as a similarity measure between the vertices x_i and x_j. A natural kernel acting on functions on V can be defined by normalization of the weight matrix as follows:

$$P(i, j) = \frac{W(i, j)}{\sum_{k \in V} W(i, k)}, \tag{1}$$

so that $\sum_{j \in V} P(i, j) = 1$. Note that P is asymmetric after the normalization. For any label vector or probability distribution, f, the multiplication Pf is a local averaging operation, with locality measured by the similarities W. Multiplication by P can also be understood as a generalization of Parzen window estimators to functions on graphs/manifolds. The operation $f'P$ (f' denotes transpose of f) can be viewed as a Markov walk of the vector f. Both Pf and $f'P$ can loosely be considered as a diffusion process.

2.2. Local Similarities

Given a graph, G, we construct another graph $\mathcal{G}$: the vertices of $\mathcal{G}$ are the same as in G, and weighted edges are those nearby ones only. In other words, those similarities between non-neighboring points (in terms of the pairwise similarity values) are set to zero. Essentially we make the assumption that local similarities (high values) are more reliable

than far-away ones; and accordingly local similarities can be propagated to non-local points through a diffusion process on the graph. This is a mild assumption widely adopted by other manifold learning algorithms [29, 21]. If ρ is a distance metric defined on the graph, then the similarity matrix can be constructed as follows:

$$W(i,j) = h(\frac{\rho(x_i, x_j)^2}{\mu\sigma^2}), \tag{2}$$

for some function h with exponential decay at infinity. A common choice is $h(x) = exp(-x)$. Note that μ and σ are hyper-parameters. σ is learned by the mean distance to K-nearest neighborhoods [37]. Using K nearest neighbor (KNN) to measure local affinity, we construct $\mathcal{G}$ with associated similarity matrix:

$$\mathcal{W}(i,j) = \begin{cases} W(i,j) & \text{if } x_j \in KNN(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Then the corresponding kernel becomes:

$$\mathcal{P}(i,j) = \frac{\mathcal{W}(i,j)}{\sum_{x_k \in KNN(x_i)} \mathcal{W}(i,k)} \tag{4}$$

Note that P carries the full information about the similarity of each data point to all others whereas $\mathcal{P}$ only encodes the similarity to nearby data points. For clarity, we call P the status matrix and $\mathcal{P}$ the kernel matrix. For the remainder of this paper, our algorithm always starts from P as the initial status and we use $\mathcal{P}$ as the kernel matrix in the diffusion process for computational efficiency.

2.3. Cross Diffusion Process with $m = 2$ Similarity Measures

Our basic system deals with input as $m = 2$ similarity measures or $m = 2$ sets of features (we will see extension to $m > 2$ later) for a set of n data samples. Let $x_i^{(j)} \in R^{d_j}$ be the features for the i-th sample in the j-th view where d_j is the dimension of the feature space for view j, then vector $x_i \triangleq (x_i^{(1)}, \ldots, x_i^{(m)})$ represents the concatenated features for the i-th data sample, and $X^{(j)} \triangleq (x_1^{(j)}, \ldots, x_n^{(j)})$ represents all samples from the j-th view. Therefore, we can obtain similarity matrix $W^{(j)}$ and $\mathcal{W}^{(j)}$ using eqn. (3). If no explicit features are given, our algorithm directly takes $W^{(j)}$ and computes $\mathcal{W}$ accordingly. Finally, $P^{(j)}$ and $\mathcal{P}^{(j)}$ are obtained by eqn. (1) and eqn. (4) respectively.

To explore the idea of mutual improvement and inspired by the co-training [4] algorithm, we introduce our proposed method, "Cross-Diffusion Process". First, we calculate the status matrices $P^{(1)}$ and $P^{(2)}$ as in eqn. (1) from two input similarity matrices; then the kernel matrices $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are obtained as in eqn. (4). Let $P_0^{(1)} = P^{(1)}$ and $P_0^{(2)} = P^{(2)}$. The cross-diffusion process is defined as:

$$P_{t+1}^{(1)} = \mathcal{P}^{(1)} \times (P_t^{(2)}) \times (\mathcal{P}^{(1)})' \tag{5}$$

$$P_{t+1}^{(2)} = \mathcal{P}^{(2)} \times (P_t^{(1)}) \times (\mathcal{P}^{(2)})' \tag{6}$$

where $P_t^{(1)}$ is the status matrix after t times' iterations. This process exchanges the status matrices each time and generates two parallel inter-changing diffusion processes. After t steps, the overall status matrix is computed as $P^{(c)} = \frac{1}{2}(P_t^{(1)} + P_t^{(2)})$. Since $\mathcal{P}$ is a KNN graph of P which can reduce some noise between instances, our cross diffusion process is robust to the noise of similarity measures.

The input of our algorithms can be raw feature vectors, pairwise distances, or pairwise similarities. We refer to this method as "**CrDP**" in the remainder of this paper. The learned status matrix $P^{(c)}$ from the above method is of different use for many machine learning tasks, such as retrieval, clustering, and classification; in this paper, we focus on the tasks of image retrieval.

Convergence Analysis

Next we prove that the two status matrices in our cross diffusion process in eqn. (5) converge. First, we need to define the distance between two kernels. Given two status matrices P and Q, the direct L1 distance can be computed as:

$$d(P\|Q) = \frac{1}{n^2} \sum_{i=1..n, j=1..n} |P(i,j) - Q(i,j)| \tag{7}$$

Theorem 1 *The distance between two status matrices $P_t^{(1)} \in \mathrm{R}^{n\times n}$ and $P_t^{(2)} \in \mathrm{R}^{n\times n}$, defined in eqn. (7) converges with $t \to \infty$.*

Proof: Based on eqn. (5), without loss of generality at, $2t+1$, we have

$$P_{2t+1}^{(1)} = \mathcal{P}^{(1)} \times (\mathcal{P}^{(2)}\mathcal{P}^{(1)})^t \times P_t^{(2)} \times ((\mathcal{P}^{(2)}\mathcal{P}^{(1)})')^t \times (\mathcal{P}^{(1)})' \tag{8}$$

$$P_{2t+1}^{(2)} = (\mathcal{P}^{(2)}\mathcal{P}^{(1)})^t \times \mathcal{P}^{(2)} \times P_t^{(1)} \times (\mathcal{P}^{(2)})' \times ((\mathcal{P}^{(2)}\mathcal{P}^{(1)})')^t \tag{9}$$

Let $\{\lambda_1, \lambda_2, ..., \lambda_r\}$ be the eigen values of $\mathcal{P}^{(2)}\mathcal{P}^{(1)}$, and

$$\lambda_1 > |\lambda_2| \geq ... \geq |\lambda_r|.$$

A general assumption for the transition matrix $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ for being irreducible and aperiodic can be made. According to the Perron-Frobenius Theorem [20], we have

$$(\mathcal{P}^{(2)}\mathcal{P}^{(1)})^t = \mathbf{1}\pi' + O(n^{m_2-1}|\lambda_2|^t),$$

where $\mathbf{1}$ and π are the top right and left eigen vectors for $\mathcal{P}^{(2)}\mathcal{P}^{(1)}$, and m_2 is the algebraic multiplicity of λ_2. Therefore, with $t \to \infty$, we have

$$(\mathcal{P}^{(2)}\mathcal{P}^{(1)})^t Q((\mathcal{P}^{(2)}\mathcal{P}^{(1)})')^t \to B, \forall Q,$$

where B is a fixed matrix dependent on π only. Therefore,

$$d(P_{2t+1}^{(1)} | P_{2t+1}^{(2)}) \to^{t=\infty} d(B|\mathcal{P}^{(1)}B(\mathcal{P}^{(1)})') = \epsilon,$$

where ϵ is the difference between a fixed matrix B w.r.t. itself after one round of diffusion. This shows that $P_t^{(1)}$ and

$P_t^{(2)}$ converges to the same status matrix B with difference in only one round of diffusion.

One can add extra regularization as optional steps to eqn. (5) to increase the robustness of our algorithm:

$$P_{t+1}^{(1)} = \mathcal{P}^{(1)} \times (P_t^{(2)}) \times (\mathcal{P}^{(1)})' + \eta I$$
$$P_{t+1}^{(2)} = \mathcal{P}^{(2)} \times (P_t^{(1)}) \times (\mathcal{P}^{(2)})' + \eta I \quad (10)$$

The reason of adding ηI is at least two-fold: (1) to avoid the loss of the self-similarity through the diffusion process; (2)to ensure more robust mass distributed. The convergence after adding the regularized term is consistency observed in various experiments (see Sec. 3). It is noted that although using these optional steps slightly improves the algorithm performance over eqn. (5), improvement of CrDP is consistently observed over the baseline and contemporary methods with or without these options.

2.4. Probabilistic Interpretation

We demonstrate the idea of cross diffusion from a probabilistic view. Given the status matrix $P_t^{(1)}$, we can define the *diffusion distance* [5] at time t as follows:

$$D_t^{(1)}(i,j) = \| P_t^{(1)}(i,:) - P_t^{(1)}(j,:) \| \quad (11)$$

This means the diffusion process maps the data space into a n-dimensional space $\Re_t^{(1)}$ in which each data point is represented by its probability to the other data points. It is reasonable to assume that for each data $\mathbf{x}_t^{(1)} \in \Re_t^{(1)}$, we have $p(\mathbf{x}_t^{(1)}) = \mathcal{N}(\mathbf{x}_t^{(1)}|\mu_t, P_t^{(1)})$. Note in the cross diffusion process, two different kernels start merge in some sense. To do this, we design an linear operator $\mathcal{P}^{(2)}$:

$$\mathbf{x}_{t+1}^{(2)} = \mathcal{P}^{(2)}\mathbf{x}_t^{(1)} + \sqrt{\eta}\varepsilon \quad (12)$$

where ε is white noise, i.e. $p(\varepsilon) = \mathcal{N}(\varepsilon|0,1)$. Under this linear operation, we have:

$$p(\mathbf{x}_{t+1}^{(2)}|\mathbf{x}_t^{(1)}) = \mathcal{N}(\mathbf{x}_{t+1}^{(2)}|\mathcal{P}^{(2)}\mathbf{x}_t^{(1)}, \eta I). \quad (13)$$

The marginal distribution of $\mathbf{x}_{t+1}^{(2)}$ is

$$p(\mathbf{x}_{t+1}^{(2)}) = \int_{\Re_t^{(1)}} \mathcal{N}(\mathbf{x}_t^{(1)}|\mathbf{x}_t^{(1)}, P_t^1)\mathcal{N}(\mathbf{x}_{t+1}^{(2)}|\mathcal{P}^{(2)}\mu_t, \eta I)d\mathbf{x}_t^{(1)}$$
$$= \mathcal{N}(\mathbf{x}_{t+1}^{(2)}|\mathcal{P}^{(2)}\mu_t, \mathcal{P}^{(2)}P_t^{(1)}(\mathcal{P}^{(2)})' + \eta I)$$

From the above equation and Eqn.(5), we can see that, the essence of cross diffusion is to do linear operation on diffusion space iteratively. Let us look at the linear operator in Eqn.(12). $\mathcal{P}^{(2)}$ is a sparse version of $P_0^{(2)}$ and only KNN information in the space $\Re_0^{(2)}$ is kept in the operator $\mathcal{P}^{(2)}$:

$$\mathbf{x}_{t+1}^{(2)}(i) = \sum_{j \in KNN^{(2)}(i)} P_0^{(2)}(i,j)\mathbf{x}_t^{(1)}(j) + \sqrt{\eta}\varepsilon$$

This projection combines information from two views. Note $\mathbf{x}_t$ is a point in the diffusion space. Instead of linear projection in the original data space, we do projection on diffusion space. The advantages of projection on diffusion space are two-folds: 1) The projection is robust to noise and scales of data points; 2)The projection incorporates the intrinsic structure of similarity manifold of the whole data set.

2.5. Geometrical Interpretation

In this section, we provide a geometrical explanation to our method. We suppose that the K-nearest-neighbors is good to measure local affinity, i.e., for any example x: (1) for some small ϵ, there are at least $K(1-\epsilon)$ same-class (or same cluster) examples in the K nearest neighbors of x; (2) if there is another class example $x^{'}$ in the K nearest neighbors of x, then $x^{'}$ does not belong to the K nearest neighbors of any other example, which means that $x^{'}$ may be an outlier. Let $pur(\pi_\theta)$ denote the purity of the connected component π_θ in graph $\mathcal{P}$, then

$$pur(\pi_\theta) = \max\Big[|\{x : x \in \pi_\theta \wedge c(x) = 1\}|/$$
$$|\pi_\theta|, |\{x : x \in \pi_\theta \wedge c(x) = -1\}|/|\pi_\theta|\Big], \quad (14)$$

where $c(x)$ denote the ground class (cluster) label of x. If $pur(\pi_\theta) \geq 1-\epsilon$ for all $1 \leq \theta \leq \lambda$ where λ is the number of the connected component in graph $\mathcal{P}$, we say that $\mathcal{P}$ is an ϵ-good graph [35]. For an easier denotation, we use the two-class case, positive vs. negative, for discussion and also focus on two views. We call a connected component as positive (negative) component if most examples in this component are positive (negative). If the K-nearest-neighbors is good to measure local affinity, obviously $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are ϵ-good graphs.

For the convenience of discussion, we let $P_1^{(1)} = \mathcal{P}^{(1)}$ and $P_1^{(2)} = \mathcal{P}^{(2)}$. With Eq.(5), we find that

$$P_{2t+1}^{(1)} \propto \left(\mathcal{P}^{(1)}\mathcal{P}^{(2)}\right)^t \mathcal{P}^{(2)} \left((\mathcal{P}^{(2)})'(\mathcal{P}^{(1)})'\right)^t \quad (15)$$
$$P_{2t+1}^{(2)} \propto \left(\mathcal{P}^{(2)}\mathcal{P}^{(1)}\right)^t \mathcal{P}^{(1)} \left((\mathcal{P}^{(1)})'(\mathcal{P}^{(2)})'\right)^t. \quad (16)$$

Theorem 2 *If the K-nearest-neighbors is good to measure local affinity, $P_{2t+1}^{(1)}$ and $P_{2t+1}^{(2)}$ are ϵ-good graphs. The number of connected components in graph $P_{2t+1}^{(1)}$ is the same as that in graph $P_{2t+1}^{(2)}$, which is no larger than that in graphs $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$.*

The proof is given in the appendix.

2.6. Extension to $m > 2$

In this section, we extend our algorithm given multiple ($m > 2$) similarities. The main idea is the same as in the

case of $m = 2$, and we just need to adjust eqn. (5) to

$$P_{t+1}^{(i)} = \mathcal{P}^{(i)} \times \left(\frac{1}{m-1}\sum_{j\neq i} P_t^{(j)}\right) \times (\mathcal{P}^{(i)})' + \eta I, \quad (17)$$

where $i = 1, \ldots, m$. The corresponding final status matrix is computed as $P^{(c)} = \frac{1}{m}\sum_{i=1}^{m} P_t^{(i)}$.

3. Experimental results

The proposed method benefits from its robustness to the parameter settings. In all of the experiments, we use the same set of parameters, which are $\mu = 0.36$ in (2), $K = 20$ in (3), $\eta = 1$ in (10). We use the term "baseline" to represent the accuracy of the initial metric without any learning techniques.

3.1. Retrieval

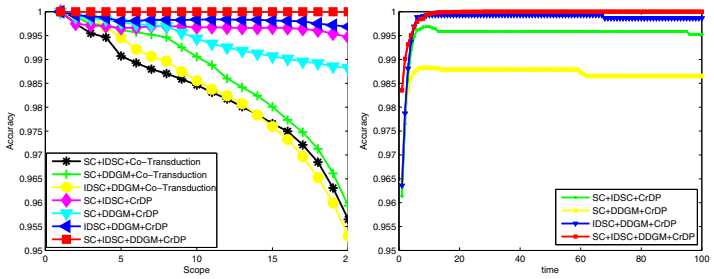
3.1.1 Shape Retrieval

The proposed algorithm is first tested for shape retrieval on a commonly used MPEG-7 database [11]. The dataset contains $1,400$ silhouette images from 70 classes, where each class has 20 different shapes. Traditionally, the performance is measured by the bull's eye score: every shape in the database is treated as a query and the accuracy of a retrieval window of size 40 is accumulated and reported. As the bull's eye score saturates with our algorithms, it is replaced by another measure called direct accuracy score in this paper: instead of using a retrieval window of size 40, we use a size 20 window. Note it is a much stricter measure than the bull's eye score: a 100% direct accuracy score means perfect retrieval results.

Two different shape matching algorithms are being fused: Shape Contexts (SC) [3] and Inner Distance (IDSC) [13], each of which outputs a $1,400 \times 1,400$ similarity matrix on the MPEG-7 shape dataset. Recent state-of-the-art multi-view methods on this dataset include a co-transduction(Co-T) fusion approach [2]. We also test the diffusion maps (DM) [5] and Bayesian Co-training(B-Co) [38] on combined descriptors. As shown in Table. 1, the advantage of our method over the competing approaches, including a direct fusion of the input metrics, is evident.

In addition, we tested the performance of the proposed method in the situation with multiple ($m > 2$) input similarities. We use another similarity computed by data-driven generative model (DDGM) [30]. Together with the other two descriptors SC and IDSC, further improvement is observed in Table. 1, which justifies the generalization capability of our method in cases with $m > 2$.

We report the performance of our algorithm and compare it to some other state-of-art methods in this dataset (see Table.1); as we can see, our method improves the baseline and the competing methods by a large margin. Our method achieves the perfect results (100% in pure accuracy which is the first time in the literature) when three descriptors are combined. Note that the baseline of the simple sum of three similarities is even worse than the baseline of combination of SC and IDSC. This suggests that a direct combination of all the similarities is not necessarily improving the overall performance. Our cross diffusion process instead uses a dynamic process to fuse the multiple metrics to make use of their complementariness.



(a) (b)

Figure 2. (a) The Accuracy-Scope curve for comparison with Co-Transduction and CrDP on MPEG-7. (b) Accuracy of CrDP on MPEG-7 over a long range of number of iterations

To fully compare to the co-transduction method which is most similar to our method, we show the Accuracy-Scope curve in Fig. (2.a). Compared with co-transduction, our method is stable over long t (as shown in Fig. (2.b)) while the performance of co-transduction drops fast when the number of iterations becomes large [2].

3.2. Face Retrieval

When only a single modality of features is available, our approach is still applicable so long as various metrics can be defined on the features. With one modality, we can still calculate different kinds of distances which provide different aspects to describe the data points (a detailed description of various distance functions can be found [1]). For example, one can use L_2 distance, L_1 distance, $L_{0.5}$ distance [1]induced metrics and cosine similarity metric [2] on the same set of features. Here, we illustrate this point by some experiments on a face retrieval application using the $AT\&T$ face image dataset [22]. This dataset consists of 400 images of 40 subjects, each one has 10 images with slight variation in facial expression and illumination. We use a very common face descriptor, LBP [18].

Retrieval accuracy is defined as the direct accuracy score with a retrieval window of size 10. The retrieval results are shown in Table.2. As is shown in the table, our method

[1]The L_p distance for n-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ is $(\sum_i^n(\mathbf{a}_i - \mathbf{b}_i)^p)^{\frac{1}{p}}$

[2]The cosine similarity for vectors a and b is $\frac{a*b}{\|a\|\|b\|}$

Table 1. **Direct accuracy scores** (see the text for definition) on MPEG-7 dataset. SC refers to [3]; IDSC refers to [13]; DDGM refers to [30]; DM refers to diffusion maps [5]; Co-T refers to [2]. B-Co refers to [38]. The baseline retrieval accuracy of each descriptor (SC,IDSC,DDGM) is 79.83, 76.07, 75.31, respectively

	SC+IDSC	SC+DDGM	DDGM+IDSC	SC+IDSC+DDGM
Baseline	88.79	84.85	83.14	87.68
DM	89.00	89.10	88.59	92.99
Co-T	94.86	94.86	95.12	95.24
B-Co	87.95	88.70	87.89	90.42
CrDP	**99.86**	**98.83**	**99.69**	**100**

Table 2. Retrieval accuracies on the $AT\&T$ face dataset. The baseline of each distance measure(L_2,L_1, $L_{0.5}$, $Cos.$) is 78.57, 78.64, 78.62, 79.80, respectively

Method	L_2+L_1	$L_2+L_{0.5}$	L_1+ $L_{0.5}$	L_2+ $Cos.$	L_1+ $Cos.$	$L_{0.5}+$ $Cos.$	L_2+L_1 $+L_{0.5}$	L_2+L_1 $+Cos.$	$L_2+L_{0.5}$ $+Cos.$	$L_1+L_{0.5}$ $+Cos.$	L_2+L_1+ $L_{0.5}+Cos.$
Baseline	81.78	81.80	81.75	82.08	82.17	82.09	82.20	82.29	82.30	82.28	83.32
DM	82.84	82.80	82.89	83.04	82.97	83.10	83.09	83.15	83.14	83.19	83.38
Co-T	84.44	84.38	84.29	84.74	84.61	84.59	84.88	84.87	84.80	84.86	84.97
B-Co	81.89	81.90	81.94	82.07	82.14	82.10	82.29	82.31	82.37	83.29	83.42
CrDP	92.38	92.65	92.49	92.58	92.67	92.70	93.25	93.41	93.35	93.38	**94.27**

has more than 8% improvement over the other existing metric fusion methods, and more than 12% improvement over the baseline with a single modality of features. We can see that other methods can just achieve a slight improvement over the baseline. Here the cross diffusion process benefits from the complementariness of two/multiple different metrics rather than different feature modalities. Since L_p distance captures the magnitude of the difference vector while cosine similarity measures the angle between two vectors, they provide complementary views which can be effectively combined by our algorithm.

These experiments show that our algorithms are effective to exploit the complementariness of two/multiple views, which can come from either different feature modalities or diverse similarity metrics defined on a single modality.

3.3. Caltech-101

Figure 3. Some sample images from the subset of the Caltech101 dataset [8] we used. They are chosen due to the relatively large number of available images within the category.

We also tested our algorithm on a well-known Caltech-101 dataset[8] which consists 101 classes and a collection of background images. We selected 12 classes (including animals, faces, building, etc.) from Caltech-101, which contains total 2788 images. These classes are chosen due to the relatively large number of available images within the category. The number of images per category varies from 41 to 800, most of which are medium resolution, i.e. about

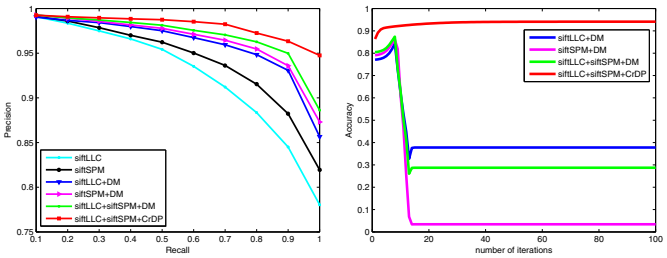
Table 3. Retrieval accuracies on the Caltech-101 subset. The best accuracy is 100. The accuracy of each descriptor(siftLLC, siftSPM) is 78.57, 80.10, respectively

siftLLC siftSPM (Baseline)	siftLLC siftSPM +DM	siftLLC+ siftSPM +Co-T	siftLLC+ siftSPM +B-Co	siftLLC+ siftSPM +CrDP
82.38	87.41	86.85	85.79	**94.11**

300×200 pixels. Fig.3 shows some samples of the subset. We use two kinds of variants of SIFT feature: SIFT with locality-constrained linear coding (siftLLC) [33] and SIFT with Spatial Pyramid Matching (siftSPM) [12]. The SIFT features are both extracted from 16×16 pixel patches on a grid with step size of 8 pixels. The codebook are obtained by standard K-means clustering with the codebook size 2,048. The distance between two images is obtained by the χ^2 distance between two feature vectors. Note that the retrieval window size is just the number of images in each category. The final accuracy is the average accuracy of each image.

The accuracy results are shown in Table.3. The proposed method improves the retrieval accuracy of baselines by about 15%, and achieved large improvement over the other methods. Note that diffusion maps(DM) is related to the proposed method and performs quite well in this dataset. This suggests that diffusion-based methods can somehow discover the intrinsic structure of natural images collection. We give a thorough comparison to diffusion maps by measuring the retrieval performance using the Precision-Recall curve (see Fig. (4)). Note that we only use precision-recall curve in Caltech101 because the number of images in each category is different. Precision-recall is a more accurate way to describe the effectiveness of the method. The other three data sets, the numbers of images in each category

are the same, so it is enough to use accuracy only. The Precision-Recall curves show that our method outperforms diffusion maps; our method is also robust to large variance in the number of images in each class. In addition, one drawback of diffusion maps is that its performance is sensitive to choice of the number of iterations (as shown in Fig. (4.b). When DM iterates too many rounds, the performance drops significantly. However, our method converges to a promising result as the number of iterations increases.



(a) (b)

Figure 4. Comparison with single-view features with diffusion maps and cross-diffusion process. (a) shows the precision and recall. (b) shows the performance against the number of iterations.

In this experiment, we test the sensitivity of the three parameters(μ in (2), K in (3), η in (10)) in the proposed method. We vary one parameters at a time while fixing the others. For instance, when we test the sensitivity of parameter μ, we try different values of μ within a range of $[0.1, 4]$ while the other two parameters K and η are fixed to be 20 and 1 respectively. The corresponding retrieval results are shown in Fig.(5). We can see that, our method is very insensitive to those three parameters. That is why in all experiments, we just use one fixed set of parameters.

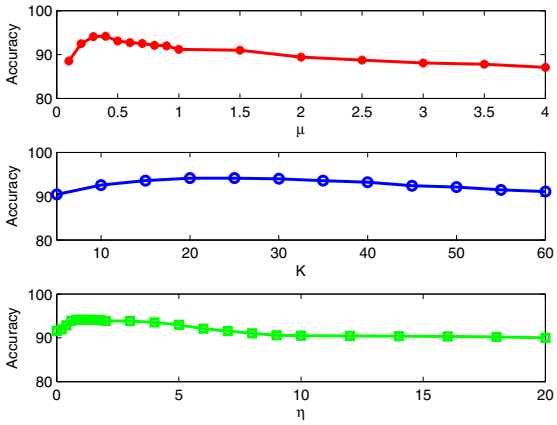


Figure 5. Sensitivity test on Caltech 101 over the set of parameters (μ, K, η) in the proposed method.

Table 4. Retrieval accuracies on the N-S data set. The best accuracy is 4. The baseline of each descriptor (GIST, SIFT) is 2.94,3.22, respectively

GIST+SIFT (Baseline)	GIST+SIFT Co-T	GIST+SIFT DM	GIST+SIFT B-Co	GIST+SIFT CrDP
3.19	3.42	3.35	3.24	**3.68**

Figure 6. The first 4 retrieved shapes in N-S dataset by GIST and SIFT. The one left the black line shows the query image. The last column shows the corresponding method.

3.4. Natural Image Retrieval

In this section, we demonstrate the performance of the proposed approach for natural image retrieval. We select the Nister and Stewenius (N-S) dataset composed of $10,200$ images [26]. The N-S dataset consists of $2,550$ objects or scenes, and each is imaged from 4 different viewpoints. Hence we have total of $2,550$ image classes, and each class has only 4 images. This is a very challenging dataset, especially for unsupervised manifold learning. We use two different image descriptors: SIFT [14] and GIST [19]. We calculate Chi-Square distance between image descriptors.

The retrieval results are shown in Table. 4. The retrieval rate is measured by the average number of correct images among the first four image returned. Therefore, the best accuracy is 4 and the higher the value the better is the result. We can see that our methods can greatly improve the baseline methods, and show better performance than the other unsupervised metric fusion methods. Fig.(6) shows some examples of retrieval results.

4. Conclusion

We have presented a cross-diffusion approach for enhancing similarity measures given two/multiple metrics. Our method takes advantage of multiple input similarity metrics and fuses them in a dynamic process. Our algorithm is easy to implement and generally applicable to a wide range of applications. The learned/fused metric can greatly improve the results of shape, face, and natural image retrieval. Significantly improvement over the-state-of-the-art has been observed on various benchmark datasets.

Acknowledgment: This work is supported by Office of Naval Research Award N000140910099 and NSF CAREER award IIS-0844566. It is also supported by NSFC

(61073097, 61021062) and 973 Program (2010CB327903).

References

[1] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Lecture Notes in Computer Science*, pages 420–434. Springer, 2001.

[2] X. Bai, B. Wang, X. Wang, W. Liu, and Z. Tu. Co-transduction for shape retrieval. In *Proc. of ECCV*, 2010.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on PAMI*, 24(4):509–522, April 2002.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. of COLT*, 1998.

[5] R. Coifman and S. Lafon. Diffusion maps. *Applied and Comp. Harmonic Ana.*, 2006.

[6] S. Dasgupta, M. L. Littman, and D. McAllester. Pac generalization bounds for co-training. In *Proc. of NIPS*, 1999.

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proc. of ICML*, pages 513–520, 2007.

[8] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:594–611, April 2006.

[9] S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In *Proc. of ICML*, 2000.

[10] A. Kumar and H. D. III. A co-training approach for multi-view spectral clustering. In *ICML*, pages 393–400, 2011.

[11] L. J. Latecki and R. Lakamper. Shape similarity measure based on correspondence of visual parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1185–1190, 2000.

[12] S. Lazebnik and C. Schmid. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of CVPR*, 2006.

[13] H. Ling and D. Jacobs. Shape classification using the inner-distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):286–299, 2007.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. of Comp. Vis.*, 60(2):91–110, 2004.

[15] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Science*, 2(1):49–55, 1936.

[16] M. Maila and J. Shi. Random walk view of segmentation, and learning spectral graph partitioning: Learning segmentation with random walk. In *Proc. NIPS*, 2001.

[17] H. B. Mitchell. *Multi-sensor Data Fusion?An Introduction*. Springer-Verlag, Berlin, 2007.

[18] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, 2002.

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[20] O. Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263, 1907.

[21] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[22] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *Proc. of 2nd IEEE Workshop of Applications of Computer Vision*, 1994.

[23] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *Proc. of NIPS*, 2003.

[24] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[25] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proc. of Workshop on Learning with Multiple Views*, 2005.

[26] H. Stewenius and D. Nister. Object recognition benchmark.

[27] S. Sun and J. Shawe-Talyor. Sparse semi-supervised learning using conjugate functions. *Journal of Machine Learning Research*, pages 2423–2455, 2010.

[28] M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *Proceedings of NIPS*, 2009.

[29] J. B. Tenenbaum, V. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[30] Z. Tu and A. Yuille. Shape matching and recognition:using generative models and informative features. In *Proc. of ECCV*, 2004.

[31] V. Vindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML*, 2005.

[32] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *TKDE*, 20(1):55 –67, 2008.

[33] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Proc. of CVPR*, 2010.

[34] W. Wang and Z.-H. Zhou. Analyzing co-training style algorithms. In *Proc. of ECML*, 2007.

[35] W. Wang and Z.-H. Zhou. A new analysis of co-training. In *ICML*, pages 1135–1142, 2010.

[36] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell. Distance metric learning with application to clustering with side-information. In *Proc. of NIPS*, pages 505–512, 2002.

[37] X. Yang, X. Bai, L. Latecki, and Z. Tu. Improving shape retrieval by learning graph transduction. In *Proc. of ECCV*, 2008.

[38] S. Yu, B. Krishnapuram, R. Rosales, H. Steck, and B. Rao. Bayesian co-training. In *Proceedings of NIPS*, 2008.

[39] Z.-H. Zhou and M. Li. Tri-training: Exploit unlabeled data using three classifiers. *IEEE Trans. Knowledge and Data Engineering*, 17(11):1529–1541, 2005.

[40] X. Zhu. Semi-supervised learning with graphs. In *Doctoral Dissertation, Carnegie Mellon University, CMU-LTI-05-192*, 2005.

5. Appendix

Proof of theorem 2.

For any positive (negative) example x, the negative (positive) example $x^{'}$ (if any) in the K nearest neighbors of x does not belong to the K nearest neighbors of any other negative (positive) examples, we know that the positive component in graph $\mathcal{P}^{(v)}$ will never have joint examples with the negative component in graph $\mathcal{P}^{(3-v)}$ $(v = 1, 2)$. We know that each connected component in graph $\mathcal{P}^{(1)}\mathcal{P}^{(2)}$ is composed of some connected components in graphs $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ which have joint examples. $P_{2t}^{(1)}$ is an ϵ-good graph and the number of connected components in graph $P_{2t}^{(1)}$ is no larger than that in graphs $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$. Obviously, the same result can be got for graph $P_{2t}^{(2)}$. Considering that $P^{(c)} = \frac{1}{2}(P_t^{(1)} + P_t^{(2)})$, we know that $P^{(c)}$ is an ϵ-good graph and the number of connected components in graph $P^{(c)}$ is no larger than that in graphs $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$. Though a general Markov kernel is often assumed to be irreducible, the convergence rate of Markov chain is also dependent on conductance, which measures the bottleneck (components). The real graph observes sparsity and locality, which makes the assumption of components not unreasonable.