

Activity Recognition Via Classification Constrained Diffusion Maps

Yunqian Ma¹, S.B. Damelin², O. Masoud³, and N. Papanikolopoulos³

¹ Honeywell labs, Honeywell International Inc.,
3660 Technology Drive, Minneapolis, MN 55418
yunqian.ma@honeywell.com

² University of Minnesota, Institute for Mathematics and its Applications
400 Lind Hall, 207 Church Hill, S.E Minneapolis, MN 55455
damelin@georgiasouthern.edu

³ University of Minnesota, Artificial Intelligence, Vision and Robotics Lab
Department of Computer Science and Engineering, Minneapolis, MN 55455
{masoud, npapas}@cs.umn.edu

Abstract. Applying advanced video technology to understand human activity and intent is becoming increasingly important for video surveillance. In this paper, we perform automatic activity recognition by classification of spatial temporal features from video sequence. We propose to incorporate class labels information to find optimal heating time for dimensionality reduction using diffusion via random walks. We perform experiments on real data, and compare the proposed method with existing random walk diffusion map method and dual root minimal spanning tree diffusion method. Experimental results show that our proposed method is better.

1 Introduction and Background

Recognition of human actions from video streams has recently become an active area of research with numerous applications in video surveillance, which is mostly motivated by the increasing number of video cameras deployed for video surveillance and the current inability of video operators to monitor and analyze large volumes of data. For predefined activities, many rule-based or logic based methods have been proposed. For example, in [1], the authors define a series of rules, e.g. entry violation, escort, theft whereas the results of [2] use a declarative model and a logic based approach to recognize predefined activities. Unfortunately a major drawback of pre-defined activity recognition approaches is that the rules developed for one activity typically may not be applicable for other activities. Indeed, different application domains may be interested in different activities. One of the key challenges in these later systems is the ability to model the activities of interest, as well as develop a methodology that allows automatic recognition of activities. In [3,4,5], the authors show that same or similar activity video sequences are clustered close to each other and far from different activity video sequences. This paper targets an automatic activity recognition

system which initially has an activity gallery that may be empty or may contain a number of initial simple activities. The system is trained by example, where input video sequences are manually labeled and the system extracts features and automatically learns the new activity.

We suppose that we are given a set of labeled video sequences as training data. The class label denotes a number of activities. Each video sequence is represented by a high dimensional feature considered isometric to a point in a high dimensional vector space. One may think that high dimension here should be an obstacle for any efficient processing of our data. Indeed, many machine learning algorithms have a computational complexity that grows exponentially with dimension. Dimensionality reduction is a way to find an isometric mapping of each video sequence into a corresponding point in Euclidean space of lower dimension where its description is considered simpler.

High dimensional spatial temporal features are associated with the nodes of a graph with a natural metric. After dimensionality reduction, a classifier (e.g. k nearest neighbor classifier) is performed on the reduced features. Using dimensionality reduction in the application of activity recognition can be found in Zhong et al. and Porikli et al. [6,7]. For example, in [6] the authors calculate the co-occurrence matrix between features, and solve for the smallest eigenvectors to find an embedding space.

In this paper, we propose a new dimensionality reduction method. The idea is to incorporate class labels information in the training data to find the optimal heating time t for dimensionality reduction using diffusion via random walks. For each heating time t , it associates a map which takes high dimensional feature to a reduced feature points. With the class labels, we perform cross validation method on the training data and then select the optimal t value which yields the smallest cross validation value. For this optimal t , we perform diffusion dimensionality reduction on the high dimensional spatial temporal feature and then use a k nearest neighbor classifier on the reduced space. We use our methods on real data, and compare the proposed method with existing random walk diffusion and dual root minimal spanning tree diffusion.

The remainder of this paper is organized as follows. In Section 2, we first describe spatial temporal features and then describe existing diffusion map methods. Section 3 describes our proposed classification constrained diffusion map method. In Section 4, we present experimental results and finally in Section 5, we present a summary.

2 Existing Diffusion Map Methods

Before we describe the existing diffusion map methods, let's briefly talk about the high dimensional Spatial temporal features used in this paper.

Davis and Bobick [10] used recursive filtering to construct feature images that represent motion: recent motion is represented as brighter than older motion. We use a similar approach described in [12]. Actions can be complex and repetitive making it difficult to capture motion details in one feature image. In this method, a weighted average at time $i \geq 1$, M_i is computed as $M_i = \alpha I_{i-1} + (1 - \alpha)M_{i-1}$,

where I_i is the image at time i and $0 \leq \alpha \leq 1$ is a fixed scalar. The feature image at time i , which we denote by F_i is computed as $F_i = |I_i - M_i|$. Note that it is the contrast of the gray level of the moving object which determines the magnitude of the feature image not the actual gray level value. To form a spatial

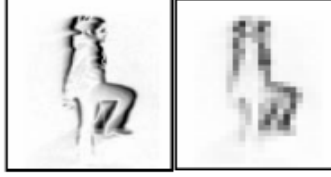


Fig. 1. An original size and a reduced size feature image from a an action of marching soldiers

temporal features, we can combine L frames together, e.g. $L = 12$ in this paper. So each spatial temporal feature is isometric to a point in $\mathbb{R}^{25 \times 31 \times 12}$, where spatial resolution can be reduced to 25×31 pixels [12], as shown in Figure 1.

Next, we describe Diffusion via Random Walks [9]. We denote X is the space of spatial temporal features in \mathbb{R}^d . First introduced in the context of manifold learning, eigenmap techniques [8,9,11] are methods to isometrically embed points of X into a lower dimensional Euclidean space. Spectral methods take into account local distortion of data points in X . The diffusion map methods belong to the spectral methods.

The existing diffusion map methods of Diffusion via random walk [9] is that it constructs a graph on X where each point is considered a node and every two nodes are connected by an edge via a non negative, symmetric, positive definite kernel $w : X \times X \rightarrow \mathbb{R}$. For example, the heat kernel can be

$$w_\sigma(\mathbf{x}_i, \mathbf{x}_j) := \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \mathbf{x}_i, \mathbf{x}_j \in X, i, j = 1, \dots, n \quad (1)$$

where σ is a kernel width parameter. The parameter σ gives the rate at which the similarity between two points decays. The weight w reflects the degree of similarity or interaction between the points $\mathbf{x}_i, \mathbf{x}_j \in X$ and depends only on the distance between \mathbf{x}_i and \mathbf{x}_j in X . Here, $\|\cdot\|$ is the Euclidean norm in \mathbb{R}^d .

A Markov chain is defined on X as follows. Given a node $\mathbf{x}_i \in X$, we define the degree of \mathbf{x}_i by $d(\mathbf{x}_i) = \sum_{\mathbf{x}_j \in X} w_\sigma(\mathbf{x}_i, \mathbf{x}_j)$. We then form a $n \times n$ affinity matrix P with entries $p(\mathbf{x}_i, \mathbf{x}_j) = \frac{w_\sigma(\mathbf{x}_i, \mathbf{x}_j)}{d(\mathbf{x}_i)}$, $i, j = 1, \dots, n$. Because $\sum_{\mathbf{x}_j \in X} p(\mathbf{x}_i, \mathbf{x}_j) = 1$, P is a transition matrix of a Markov chain on the graph of the members of X . Taking powers of P in steps $t \geq 1$, produces probability functions $p_t(\mathbf{x}_i, \mathbf{x}_j)$ which measure the probability of transition from \mathbf{x}_i to \mathbf{x}_j in t steps. Since w_σ is symmetric, P has a sequence of n eigenvalues

$$1 \geq \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$$

and a collection of $1 \leq d_r \leq n$ right eigenvectors $\{\phi_{d_r}\}$ so that for each fixed $t \geq 1$,

$$P^t \phi_{d_r} = \lambda_{d_r}^t \phi_{d_r}.$$

Each eigenvector is a signal over the data points and the eigenvectors form a new set of coordinates on X . For any choice of t , the mapping

$$\Psi_t : \mathbf{x}_i \rightarrow (\lambda_1^t \phi_1(\mathbf{x}_i), \dots, \lambda_{d_r}^t \phi_{d_r}(\mathbf{x}_i))^T \quad (2)$$

is an isometric embedding of X into \mathbb{R}^{d_r} and the function

$$\alpha(\mathbf{x}_i, \mathbf{x}_j) := \|\Psi(\mathbf{x}_i) - \Psi(\mathbf{x}_j)\|, \quad i, j = 1, \dots, n$$

defines a metric on the graph given by the nodes of X . Here $\|\cdot\|$ denotes the Euclidean norm in \mathbb{R}^{d_r} .

The reason that spectral clustering methods work [13] is that with sparse kernel matrices, long range affinities are accommodated through the chaining of many local interactions as opposed to standard Euclidean distance methods - e.g. correlation - that impute global influence into each pair wise affinity metric, making long range interactions wash out local interactions.

Another diffusion methods, proposed by Grikschat et al. [11], is a dual root minimal spanning tree diffusion method, we put the description in Appendix, since we also compared our proposed method with those method.

3 Proposed Method

We are given a set of labeled video sequences (\mathbf{x}_i, y_i) , $i = 1, \dots, n$ as training data and an unlabeled set of testing data \mathbf{x}_i , $i = 1, \dots, m$ where each $\mathbf{x}_i \in \mathbb{R}^d$ represents a d dimensional spatial temporal feature, and $y_i \in \{1, \dots, p\}$ is a class label in p activities.

Spatial temporal features \mathbf{x}_i , $i = 1, \dots, n + m$ are associated with the nodes of a graph with a natural metric given in section 2 for dimension reduction using diffusion via random walks. As described in section 2, for each fixed positive integer t , we have a map

$$t : \mathbf{z}_i^{(t)} = \mathbf{x}_i \rightarrow (\lambda_1^t \phi_1(\mathbf{x}_i), \dots, \lambda_{d_r}^t \phi_{d_r}(\mathbf{x}_i))^T, \quad i = 1, \dots, n + m \quad (3)$$

Here, for any t , we produce from $n + m$ points \mathbf{x}_i a new set of $n + m$ reduced points $\mathbf{z}_i^{(t)} \in \mathbb{R}^{d_r}$ in an Euclidean space of reduced dimension d_r . The parameters t introduce weights as multiplication of eigenvalues by eigenvectors.

From another point of view, we consider the matrix P^t where P is the affinity matrix defined in Section 2, which can be viewed as a transition matrix of a Markov chain on the nodes. Taking powers of P in steps t , produces probability functions which measure the probability of transition from \mathbf{x}_i to \mathbf{x}_j in t steps. (For this reason, the maps are diffusion related). In this paper, we propose to use class label information in the training data to choose an optimal t value for dimensionality reduction. The detail is as follows:

For each $t \geq 1$, we use cross validation on the training points $(\mathbf{z}_i^{(t)}, y_i), i = 1, \dots, n$. Specifically, we use leave-one-out cross validation defined by

$$CV^{(t)} := \frac{1}{n} \sum_{i=1}^n L(f^{\hat{-}i}(\mathbf{z}_i^{(t)}), y_i) \quad (4)$$

where $f^{\hat{-}i}$ is the fitting function computed with the i th part of the data removed and L is 1 if $f^{\hat{-}i}(\mathbf{z}_i^{(t)}) = y_i$ and 0 otherwise.

We then select the optimal t value t^{opt} defined as

$$t^{\text{opt}} = \text{argmin}(CV^{(t)}) \quad (5)$$

which yields the smallest cross validation value.

We perform t^{opt} -step random walk diffusion and arrive the resulting low dimensional feature $\mathbf{z}_i^{t^{\text{opt}}}$. After that, we use k nearest neighbor classifier for the multi-class classification on the reduce space's of the test data $\mathbf{z}_i^{t^{\text{opt}}}, i = 1, \dots, m$.

4 Experimental Results

In this experiment, the video sequences were recorded using a single stationary monochrome CCD camera mounted in such a way that the actions are performed parallel to the image plane. The data set consists of actions performed by $s = 29$ different people. Each person performed $p = 8$ activities, as shown in Figure 2: walk, run, skip, line-walk, hop, march, side-walk, side-skip. The location and size of the person in the image plane is assumed to be available (e.g., through tracking). Each activity sequence by each person includes a full cycle of the activity. The number of frames per sequence therefore depends on the speed of each action.

In our experiments, we used the data for eight of the 29 subjects for training (64 video sequences). This leaves a test data set of 168 video sequences performed by the remaining 21 subjects. The training instances have label. The number of selected frames was arbitrarily set to 12. So, the full dimension d of the space is 775×12 dimensions (as shown in Section 2).

For performance evaluation, we calculated prediction risk on the test data by the formula (6):

$$R_{\text{pred}} := \frac{1}{m} \sum_{i=1}^m L(\hat{y}_i, y_i). \quad (6)$$

Finally we compare our new method with the existing random walk diffusion and dual root minimal spanning tree diffusion.

Table 1 shows the results of prediction risk (error) using our proposed method. Table 2 and Table 3 are the results using existing Random walk diffusion+KNN and Dual rooted diffusion+KNN;

The best result of the classification error in Table 1 is 36.31%, which is better than the corresponding best result from Table 2 and Table 3 (42.86%, 57.14%



Fig. 2. Frames from walk, run, skip, march, line-walk, hop,side-walk,side-skip actions respectively

Table 1. Classification error using proposed method of cross validation to select best heating time

	d=3	d=5	d=20	d=50	d=100	d=150	d=200
$\sigma = 6$	53.57%	53.57%	51.79%	53.57%	57.74%	70.24 %	73.21%
$\sigma = 8$	57.14%	45.24%	44.05%	44.05%	46.43%	44.05%	50%
$\sigma = 10$	54.67%	56.55%	43.45%	41.07%	44.64%	47.02%	44.05%
$\sigma = 12$	50.6%	53.57%	38.69%	36.31%	39.88%	40.48%	41.67%
$\sigma = 14$	57.74%	50.6%	44.64%	47.02%	42.26%	42.26%	41.67%

Table 2. Classification error using using existing Random walk diffusion

	d=3	d=5	d=20	d=50	d=100	d=150	d=200
$\sigma = 6$	53.57%	53.57%	51.79%	56.55%	59.52%	73.21%	84.52%
$\sigma = 8$	58.93%	45.24%	50.6%	48.81%	59.52%	69.64%	87.5%
$\sigma = 10$	57.14%	56.55%	44.64%	44.64%	58.33%	69.05%	86.9%
$\sigma = 12$	62.5%	55.95%	42.86%	44.64%	60.12%	70.83%	87.5%
$\sigma = 14$	57.74%	50.6%	44.64%	47.02%	60.12%	72.62%	87.5%

Table 3. Classification error using Dual rooted diffusion, where $v = \max(\max(Ahop))$

	d=3	d=5	d=20	d=50	d=100	d=150	d=200
$\sigma = 1/6 * v$	64.29%	64.88%	63.69%	66.07%	65.48%	74.4%	83.33%
$\sigma = 1/8 * v$	61.9%	61.31%	63.69%	61.31%	64.88%	73.81%	88.1%
$\sigma = 1/10 * v$	61.31%	63.69%	57.14%	62.5%	66.67%	70.83%	83.93%
$\sigma = 1/12 * v$	60.71%	60.12%	60.12%	61.31%	66.07%	74.4%	82.74%
$\sigma = 1/14 * v$	59.52%	60.12%	63.1%	59.52%	64.29%	75%	82.74%

respectively). Also the results in Table 1 is stable and are considerably lower and less sensitive to the choice of σ than those of Table 2 and Table 3. The reason is that using proposed method the choice of optimal heating time t compensates for the sensitivity of σ .

We also use different k -value in the k -nearest neighbor classifier, and have the similar experimental results to the case $k=3$.

5 Summary

In this paper, we studied the problem of activity recognition via classification of spatial temporal feature actions in the video surveillance application. In practical application, these high level semantic learning performance also depends on the quality of low level processing, such as motion detection and motion tracking. Sometimes, there is bad quality of low level processing. For example, under very noisy video environment, shadows, occlusion, the current low level processing can not reach 100% satisfaction of the needs of activity recognition. We discuss the robust tracking algorithm in [14], so in this paper we assume the low level processing motion detection and motion tracking is well done.

In this paper, our new idea is to use class labels to find optimal heating times for dimension reduction using diffusion via random walks. We used our methods on real data, and compared our new method with the existing diffusion maps method for random walk diffusion and dual root minimal spanning tree diffusion. The results by our proposed method are shown to be considerably better as the choice of optimal heating time.

The activity we discuss in this paper is simple activity, in the future work we will perform research on complex activity, which is a combination of a series simple activities.

Acknowledgement. The authors want to thank Professor A. Hero from University of Michigan and Dr. T. Wittman from University of Minnesota for useful discussions.

References

1. V. D. Shet, D. Harwood and L. S. Davis, *VidMAP: Video Monitoring of Activity with Prolog*, IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), Como, Italy, September 2005.
2. V. Vu, F. Bremond and M. Thonnat, *Video surveillance: human behaviour representation and on-line recognition*, The Sixth International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Podere d'Ombriano, Crema, Italy, September 2002.
3. Y. Ma, B. Miller, P. Buddharaju and M. Bazakos, *Activity Awareness: From Pre-defined Events to New Pattern Discovery*, 2006 IEEE International Conference on Computer Vision Systems, New York, USA, January 5-7, 2006.
4. R. Chellappa, A. Chowdhury and S. Zhou, *Recognition of Humans and Their Activities Using Video*, Morgan Claypool, 2005.

5. N. Jin and F. Mokhtarian, *Human Motion Recognition based on Statistical Shape Analysis*, Proc. IEEE International Conference on Advanced Video and Signal based Surveillance, pp. 4-9, 2005.
6. H. Zhong, J. Shi and M. Visontai, *Detecting Unusual Activity in Video*, IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2004.
7. F. Porikli and T. Haga, *Event detection by Eigenvector Decomposition using object and feature frame*, CVPR workshop, pp. 114-114, 2004.
8. M. Belkin, P. Niyogi, *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*, Neural Computation, June 2003; **15** (6), pp. 1373-1396.
9. R. R. Coifman, S. Lafon, *Diffusion Maps*, submitted to Applied Computational and Harmonic Analysis, 2004.
10. J. Davis and A. Bobick, *The Representation and Recognition of Action Using Temporal Templates*, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico, pp. 928-934, June 1997.
11. S. Grikschat, J. Costa, A. Hero and O. Michel, *Dual rooted diffusions for clustering and classification on manifolds*, 2006 IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Toulouse France, 2006.
12. O. Masoud, N.P. Papanikolopoulos, *A method for human action recognition*, Image and Vision Computing, **21**, no. 8, pp. 729-743, Aug. 2003.
13. A. Ng, M. Jordan and Y. Weiss, *On spectral clustering: analysis and an algorithm*, Neural Information Processing Systems, **14**, 2001.
14. Y. Ma, Q. Yu and I. Cohen, *Multiple Hypothesis Target Tracking using Merge and Split of Graphs Nodes*, 2nd International Symposium on Visual Computing, Lake Tahoe, Nevada, Nov. 6-8, 2006.

Appendix: Diffusion Via Dual Root Minimal Spanning Trees

Starting with two random walks on different points \mathbf{x}_i and \mathbf{x}_j in X , when will two paths generated hit each other. More precisely, given $\mathbf{x}_i \in X$, we compute a greedy minimal spanning tree and define the distance d between two points \mathbf{x}_i and \mathbf{x}_j as the number of greedy iterations required so that two greedy minimal spanning trees rooted on each point \mathbf{x}_i and \mathbf{x}_j in X will intersect. We set σ to be $1/C \cdot \max(\max(Ahop))$ where $C > 1$ and $Ahop$ is the matrix of all pairwise distances - the hitting times between diffusions from different pairs of points. (In [11] C is taken as 10). This is an adaptive normalization in the sense that it makes the kernel decay on the order of $1/C$ of the maximum of the hitting times. An affinity matrix P is calculated with the weight w_σ with distance given by the hitting time between points x and y and the eigenvectors of P are used for a dimension reduction map.