Learning Multi-modal Densities on Discriminative Temporal Interaction Manifold for Group Activity Recognition

Ruonan Li, Rama Chellappa
Center for Automation Research, University of Maryland
College Park, MD 20742, USA
{liruonan,rama}@umiacs.umd.edu

Shaohua Kevin Zhou
Siemens Corporate Research
Princeton, NJ 08540, USA
shaohua.zhou@siemens.com

Abstract

While video-based activity analysis and recognition has received much attention, existing body of work mostly deals with single object/person case. Coordinated multi-object activities, or group activities, present in a variety of applications such as surveillance, sports, and biological monitoring records, etc., are the main focus of this paper. Unlike earlier attempts which model the complex spatial temporal constraints among multiple objects with a parametric Bayesian network, we propose a Discriminative Temporal Interaction Manifold (DTIM) framework as a data-driven strategy to characterize the group motion pattern without employing specific domain knowledge. In particular, we establish probability densities on the DTIM, whose element, the discriminative temporal interaction matrix, compactly describes the coordination and interaction among multiple objects in a group activity. For each class of group activity we learn a multi-modal density function on the DTIM. A Maximum a Posteriori (MAP) classifier on the manifold is then designed for recognizing new activities. Experiments on football play recognition demonstrate the effectiveness of the approach.

1. Introduction

In this work we model and recognize *coordinated group activities* involving *multiple objects* from videos. Human activity analysis and classification has been a research focus of computer vision community for over two decades [2, 18]. However, most previous work focused on single object cases, where the motion and dynamics of an individual object are investigated. Activities of multiple objects exist widely in surveillance, sports, and biological observation records, *etc.*, and consequently modeling and analysis of multi-object activities will be useful in these applications. Although the multi-object tracking problem has been extensively studied [24], much less attention has been paid to putting the tracked motion pattern of the whole group in a learning and recognition framework.

In a less complex scenario the individuals in a group undergo a structurally fixed motion [13] or follow identical dynamics or trajectories [22]. In the former [13] only a deviation from a modeled formation is detected; in the latter [22], an 'abnormal' activity is claimed when the configuration of the individuals exceeds an admissible bound. However, more meaningful semantics may be extracted for less-structured but coordinated activities. In other words, the individual objects will have distinctive and varying motion patterns but the group collectively demonstrates an underlying activity with an explicit sematic identity. A most illustrative example is a football game, in which we would like to recognize the strategy used in each play rather than individual players' movements. Similar examples/applications exist in other domains, *e.g.*, activities of a group of social insects like bees [23]. In this paper we use football play modeling and recognition as our primary example and in particular test our algorithm on football videos.

A group activity usually occurs according to a planned goal. In each football play, the offensive players will collaboratively follow a pre-determined strategy. The individual action of each player, meanwhile, is also a result of interaction with and response to the motion of other players. A specific group activity pattern, therefore, is determined by the *interactions* among objects and their *temporal evolution*. Modeling and recognizing the temporal inter-object relationship (*i.e.*, the group activity pattern) has been mostly handled using a Bayesian network framework [8, 10, 6, 17, 7]. Bayesian network formulation, though successfully applied to modeling activities of single object or motion, has drawbacks when dealing with group activities. To completely characterize the role of individual objects, their action primitives, interactions, and overall plan, the complexity of the network turns out to be prohibitively high. This inherent difficulty manifested itself in previous work (*e.g.*, [10]), where individual objects' identities, roles and their individual action primitives were pre-labeled. As si-

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

multaneous recognition of individual actions and group activity pattern is computationally intensive, the number of objects considered previously, was usually small, which is not the case for a large group as a football team. Compared to the size of the state space and feature space of the network, the amount of available training data is insufficient. Thus not only the probabilistic dependence might very possibly be 'over-fitted', but also necessary priors are hard to learn from available data.

A recent work [25] employs the causality of time-series to describe pairwise activities, but this tool is difficult to be extended to more objects. Also note that extensive work has been done in sports video analysis, especially for football or soccer [15, 16, 9]. These efforts attempted to detect or recognize specific types of semantics in the games using camera motion, color, low-level motion, field markers, lines, texture, and so on, but did not treat the plays as multi-object group activities. The approaches are less useful in areas other than sports.

The work [10], most similar to ours, designed large connected Bayesian networks for football play recognition. In contrast, we explore a 'data-driven' approach. Here we only assume that the players' roles and their motion trajectories are already available. For the former, we may recognize the roles from the initial play formation with the help of landmark shape theory [5], and for the latter we may employ a multi-object tracker [24]. These two problems are still being researched and beyond the scope of this work. Specifically, we describe a group activity pattern with a full four-dimensional object-time interaction tensor, and learn an optimized tensor reduction kernel to condense it to a discriminative temporal interaction matrix. The temporal interaction matrix serves as a compact 'descriptor' for the group activity pattern, and is empirically stable under view changes. More importantly, given a Riemannian metric the set of all temporal interaction matrices forms a Riemannian manifold, on which we are able to establish a probabilistic framework to characterize every class of group activity pattern. We call this manifold Discriminative Temporal Interaction Manifold (DTIM). To learn a multi-modal 'likelihood' density for each class, we create a basic exponential density component on the manifold, and incrementally build up the complete manifold-resided densities with the basic components. With the established framework, a MAP classifier is used to recognize a new group activity.

The rest of the paper is organized as follows. In Section 2 we obtain a view-stable and discriminative temporal interaction matrix, via an optimized tensor reduction, to compactly characterize each group activity. Then in Section 3 we focus on the space of temporal interaction matrices, *i.e.*, DTIM, and in particular create a basic probability density on this non-linear manifold by exploiting its geometric property. To account for possibly multi-modal likelihood distribution of temporal interaction matrices on DTIM, in Section 4 we introduce an incremental, or boosting procedure to build the complete model with the basic components. Finally, we show the performance using data from football plays in Section 5. See Figure 1 for a general flow chart of the proposed approach.

2. View-Stable Discriminative Temporal Interaction Matrix

As mentioned, a coordinated group activity pattern is characterized by the temporally evolving interactions among objects. To describe mathematically the interaction, we define the object-time interaction tensor as $Y(t_1, t_2, p_1, p_2)$, where $1 \leq t_1, t_2 \leq T$, $1 \leq p_1, p_2 \leq P$. Here T is the duration during which we observe the group activity, and P is the total number of objects involved in the activity (*e.g.*, the total number of players in a football play). The term $Y(t_1, t_2, p_1, p_2)$ describes the 'interaction' between object p_1 at time t_1 and object p_2 at time t_2. The interaction can be interpreted in multiple ways. Since the point trajectories for objects are assumed to be available in this work, we simply take the distance between object p_1 at time t_1 and object p_2 at time t_2 as the interaction term $Y(t_1, t_2, p_1, p_2)$. Once other features are available, they can be incorporated to provide a more complete description.

The four-dimensional tensor Y is in a sense a full but possibly redundant descriptor for the activity pattern. We now seek a more compact and discriminative descriptor, namely, a temporal interaction matrix $X(t_1, t_2)$ via a tensor reduction mapping $R : Y \mapsto X$. The motivation for this reduction is two-fold. On the one hand, as the tensor is in a high dimensional space, a dimensionality reduction is generally necessary so that the need for many training samples may be reduced. On the other hand, more importantly, the temporal interaction matrix empirically turns out to be quite stable when view changes, though it is not strictly view-invariant. Figure 2 shows this 'view-stability' of the temporal interaction matrix obtained from the discriminative tensor reduction method presented below. Figure 2(a) and 2(b) show the players' motion trajectories for the same play under different views, and Figure 2(c) and 2(d) show the corresponding temporal interaction matrices, which appear quite close to each other. Although the example is synthesized from a play diagram, the same behavior is also observed for real trajectories. The search for a discriminative and view-stable matrix descriptor is also inspired by the previous work on video self-similarity [4] and a recent one [11], in which view-stability is observed in 'self-similarity matrix' of point trajectories for a single person action.

Instead of using an arbitrary tensor reduction mapping R, here we use a $P \times P$ matrix reduction kernel A. Let $Y(t_1, t_2)$ to be the $P \times P$ matrix sliced from tensor

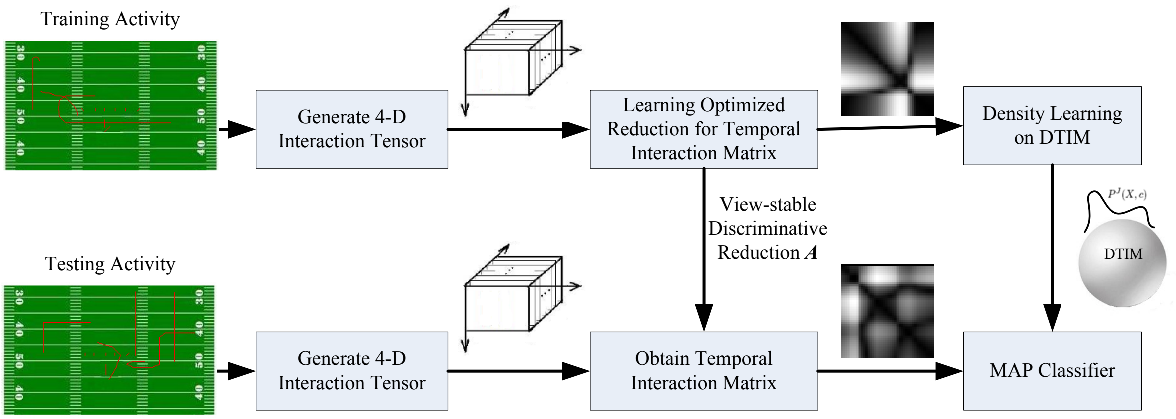
Training Activity
Generate 4-D Interaction Tensor
Learning Optimized Reduction for Temporal Interaction Matrix
Density Learning on DTIM
View-stable Discriminative Reduction A
$P^J(X,c)$
DTIM
Testing Activity
Generate 4-D Interaction Tensor
Obtain Temporal Interaction Matrix
MAP Classifier

Figure 1. The flowchart of the modeling and recognition framework.

$Y(t_1, t_2, p_1, p_2)$ when t_1, t_2 are fixed, and we define

$$x(t_1, t_2) = tr(A^T Y(t_1, t_2)) \tag{1}$$

and

$$X = R(Y) = \frac{x}{\|x\|}. \tag{2}$$

Note that both $Y(t_1, t_2)$ and A are symmetric due to the interpretation of 'interaction' above. Similar notion of symmetry also holds for X. The normalization of x helps to maintain a constant scale for X.

Note that A in fact weighs each element of $Y(t_1, t_2)$ with its corresponding element. Therefore, A serves as a pairwise interaction selector, which emphasizes the interaction between the ith and the jth objects if $A(i, j)$ is large. If we nomially take A to be the identity matrix, *i.e.*, discarding the interaction between different objects but only keeping the objects' self-motion, then the resulting temporal interaction matrix X is essentially equivalent to the one used in [11]. Therefore, the question is: is there another A (or weighting pattern) other than the identity matrix, which can achieve intra-class view-stability as well as better inter-class separability? To get such an optimized tensor reduction kernel, we mathematically enforce view-stability and separability in our optimization target as described below. In other words, we formally look for view stability instead of achieving it in an ad-hoc manner.

From now on we use different subscripts to denote temporal interaction matrices from different sample group activities. A pairwise similarity between the kth sample X_k and the lth sample X_l, $s(k, l)$, is defined accordingly as

$$s(k, l) = tr(X_k^T X_l) = \frac{< x_k, x_l >}{\|x_k\| \|x_l\|}. \tag{3}$$

Then the target function to be maximized is defined as

$$J(A) = \sum_k (\alpha\beta \sum_{l \in C_1(k)} s(k, l) + \alpha(1-\beta) \sum_{l \in C_2(k)} s(k, l) - (1-\alpha) \sum_{l \in C_3(k)} s(k, l)) \tag{4}$$

where $C_1(k)$ is the set of same activities as k but from possibly different views, $C_2(k)$ is the set of activities different from k but belonging to the same class as k, and $C_3(k)$ is the set of activities not belonging to the class of k. By maximizing $J(A)$ with respect to A with the controllable parameters $0 < \alpha, \beta < 1$, we are able to find an optimized A such that the cross-view similarity and intra-class similarity are both maximized while the inter-class similarity is minimized. In other words, the resulting A will weigh the interactions between every pair of objects properly such that view-stability and class separability are simultaneously achieved.

To perform the above maximization, we take a gradient ascent based approach due to the non-linearity of the target function with respect to A. To evaluate $\nabla_A J(A) = \frac{\partial J(A)}{\partial A}$, we evaluate $\frac{\partial s(k,l)}{\partial A}$. After some calculation it can be shown that

$$\frac{\partial s(k, l)}{\partial A} = \frac{1}{\|X_k\| \|X_l\|} (\sum_{t_1, t_2} (X_k(t_1, t_2) Y_l(t_1, t_2) + X_l(t_1, t_2) Y_k(t_1, t_2)) - b(k, l)) \tag{5}$$

where

$$b(k, l) = tr(X_k^T X_l) (\frac{\sum_{t_1, t_2} X_k(t_1, t_2) Y_k(t_1, t_2)}{\|X_k\|^2} + \frac{\sum_{t_1, t_2} X_l(t_1, t_2) Y_l(t_1, t_2)}{\|X_l\|^2}). \tag{6}$$

The optimization steps are implemented as repeated line searches along the gradient directions specified in (5) and

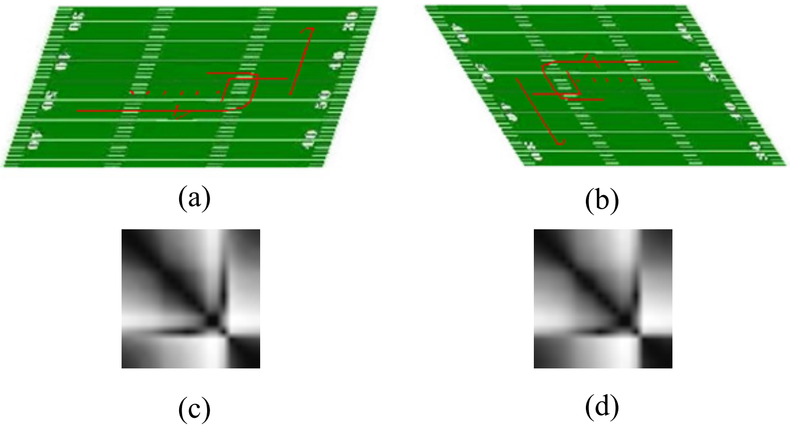

Figure 2. (a)(b):Simulated players' trajectories from different view points from play diagram of play type p51curl [10]; (c)(d): The corresponding temporal interaction matrices obtained using the view-stable optimization.

(6). Because a global maximum is not guaranteed, we initialize A from multiple symmetric matrices in multiple optimization processes and pick out the one yielding the maximum $J(A)$.

3. Basic Exponential Distribution on the Discriminative Temporal Interaction Manifold

With an optimized tensor reduction kernel A, we compute the temporal interaction matrix X_i from the full interaction tensor Y_i for the ith activity sample. As has been shown, the temporal interaction matrices serve as compact, view-stable, and discriminative descriptors for group activities, and we would like to establish a probabilistic generative model for them to characterize the activity class distribution. However, note that X is symmetric with a unit norm and thus the space of all X's is not Euclidean. Therefore, to establish the probabilistic setting we need to first exploit the geometric property of the space and then build a probability distribution on it.

3.1. The Riemannian Property of the Discriminative Temporal Interaction Manifold

Although the set of all temporal interaction matrices is not an Euclidean space, with a properly defined Riemannian metric, it becomes a Riemannian manifold. For any two elements X_1' and X_2' in the tangent space $\mathscr{T}_X$ at X, the Riemannian metric is defined as

$$< X_1', X_2' > \triangleq tr(X_1'^T X_2'). \tag{7}$$

A related case was detailed recently in [21]. This manifold is what we mentioned above as Discriminative Temporal Interaction Manifold (DTIM), denoted as $\mathscr{X}$.

With the above defined Riemannian metric the basic geometry of DTIM is straightforward. The intrinsic distance between two temporal interaction matrices X_1 and X_2 is given by

$$d(X_1, X_2) = \arccos < X_1, X_2 >, \tag{8}$$

where

$$< X_1, X_2 > \triangleq tr(X_1^T X_2). \tag{9}$$

The geodesic , *i.e.*, the curve of minimum length connecting two temporal interaction matrices X_1 and X_2 , is given by

$$X(\lambda) = \frac{(1-\lambda)X_1 + \lambda X_2}{\lambda^2 + (1-\lambda)^2 + 2\lambda(1-\lambda) < X_1, X_2 >} \tag{10}$$

where λ is a real parameter between 0 and 1.

The exponential map and logarithmic map are important for manipulations on the Riemannian manifold. For DTIM defined above, the exponential map $\mathscr{E}_X : \mathscr{T}_X \rightarrow \mathscr{X}$ for $X' \in \mathscr{T}_X$ is defined as

$$\mathscr{E}_X(X') = \cos(< X', X' >^{\frac{1}{2}})X + \frac{\sin(< X', X' >^{\frac{1}{2}})}{< X', X' >^{\frac{1}{2}}} X'. \tag{11}$$

The logarithmic map $\mathscr{L}_X : \mathscr{X} \rightarrow \mathscr{T}_X$, which is actually the inverse map of exponential map, is then given by

$$\mathscr{L}_X(X_m) = \frac{\arccos(< X, X_m >)}{< X^*, X^* >^{\frac{1}{2}}} X^* \tag{12}$$

where

$$X^* = X_m - < X, X_m > X. \tag{13}$$

It is worth noting that the temporal interaction matrix and the corresponding DTIM investigated here are closely related to and rooted in the general theoretic framework of information geometry [3, 12]. We only present the necessary geometric properties to be employed in the subsequent section. For further study on information geometry the reader is referred to [3, 12].

3.2. A Basic Exponential Distribution on DTIM and its Parameter Estimation

It is expected that the temporal interaction matrices from different activity classes will reside distinctively on the DTIM. This motivates us to establish a probabilistic approach on DTIM to model the distribution of the temporal interaction matrices.

The 'Gausssian' distribution on a Riemannian manifold is initially addressed in [19] with the help of exponential/logrithmic mapping between the manifold and the tangent plane. Here, we take a direct approach to define a uni-modal exponential distribution for the temporal interaction matrix as

$$p(X;\mu,\sigma,z) = \frac{1}{z}\exp(-\frac{d^2(X,\mu)}{2\sigma^2}), \tag{14}$$

where μ is regarded as the 'center' of temporal interaction matrices, σ characterizes the scattering of the matrices on the manifold, and z is a normalization factor. Moreover, d is the intrinsic distance defined in (8). A temporal interaction matrix intrinsically close to the center will have a high probability value.

To estimate the parameters involved in the distribution p, a statistical approach based on observed samples is practically useful. Generally, we will have a set of weighted samples $\{(X_1,w_1),(X_2,w_2),\cdots,(X_N,w_N)\}$ observed from the distribution. We define the weighted Karcher mean,

$$\mu = \arg\min_{\psi}\sum_{i=1}^{N} w_i d^2(X_i,\psi), \tag{15}$$

to be the mean parameter μ. To numerically find μ, the iterations

$$\mu'^{(g+1)} = \frac{\sum_{i=1}^{N} w_i \mathscr{L}_{\mu^{(g)}}(X_i)}{\sum_{i=1}^{N} w_i} \tag{16}$$

and

$$\mu^{(g+1)} = \mathscr{E}_{\mu^{(g)}}(\mu'^{(g+1)}). \tag{17}$$

alternate and $\mu^{(g)}$ will converge to the weighted Karcher mean as g increases. Here $\mathscr{E}$ and $\mathscr{L}$ are the exponential and logrithmic maps given in (11) and (12) respectively.

Once the mean is determined, the scattering factor can be defined in a similar manner as

$$\sigma = (\frac{\sum_{i=1}^{N} w_i d^2(X_i,\mu)}{\sum_{i=1}^{N} w_i})^{1/2}. \tag{18}$$

The calculation for the normalization factor z is analytically infeasible and we need to take the Monte Carlo approach to find the integral

$$I = \int_{\mathscr{X}} \exp(-\frac{d^2(X,\mu)}{2\sigma^2})dX \tag{19}$$

and consequently the estimate is $z = 1/I$. To perform the Monte Carlo integration we need to generate uniformly distributed samples on DTIM. To achieve this note that a $T \times T$ temporal interaction matrix is essentially equivalent to a $(1+T)T/2$ dimensional unit vector. Therefore we generate $(1+T)T/2$ dimensional homogeneous Gaussian vectors and scale them into unit length. Then we transform the unit length vectors to temporal interaction matrices, which become uniformly distributed on DTIM.

4. Learning Multi-Modal Densities on DTIM

Suppose we have a training set $\{(X_1,c_1),(X_2,c_2),\cdots,(X_M,c_M)\}$ where $c_i \in \{1,2,\cdots,C\}$ is the activity class label for the ith activity sample and there are totally C classes of group activities. Trivially we may learn a uni-modal distribution for each activity class using the method in the previous section. However, the actual scattering of temporal interaction tensors on DTIM may not be well approximated by a uni-modal model. This motivates the necessity to learn a multi-modal density for each activity class to achieve a better classification performance.

We aim to model the joint probability density function of the temporal interaction matrix X_i and the class label c_i, denoted as $P^J(X_i,c_i)$, defined as

$$P^J(X_i,c_i) = \sum_{j=1}^{J} b^j f^j(X_i,c_i) = \sum_{j=1}^{J} b^j \pi_{c_i}^j p_{c_i}^j(X_i) \tag{20}$$

where $p_{c_i}^j(X_i) = p(X_i;\mu_{c_i}^j,\sigma_{c_i}^j,z_{c_i}^j)$ is the uni-modal exponential component introduced in (14). Here we regard the joint probability as a linear mixture of J uni-modal likelihood functions where the jth component is $p_{c_i}^j(X_i)$. b^j is taken as the mixing coefficient for the jth component, which for convenience is assumed to be independent of the class labels. $\pi_{c_i}^j$ is the class prior for class c_i in the jth component. For simplicity we take $\pi_{c_i}^j \equiv \frac{1}{C}$ regardless of j or c_i. Apparently, the 'mixture-of-p' distribution will behave as the 'mixture-of-Gaussian' in an Euclidean space, providing us the capability to approximate the irregular distribution on a Riemannian manifold analytically.

The determination of a proper number of components J is non-trivial, preventing us from directly applying an Expectation-Maximization (EM) procedure to learn all the components and mixing coefficients. Instead, we build up this linearly-combined multi-modal distribution in an incremental manner. In other words, we will 'boost' the distribution on DTIM. Suppose that we have in some way achieved a J-component density $P^J(X,c)$ and we want to update it into a $(J+1)$-component one by linearly mixing it with a new $f(X,c)$,

$$P^{J+1}(X,c) = (1-\alpha)P^J(X,c) + \alpha f(X,c), \tag{21}$$

and we aim to maximize the log posteriori class probability for all samples in the training set $\sum_{i=1}^{M}\log P^{J+1}(c_i|X_i)$. Expanding the expression we get

$$\begin{aligned}\sum_{i=1}^{M}\log P^{J+1}(c_i|X_i) &= \sum_{i=1}^{M}\log\frac{P^{J+1}(X_i,c_i)}{P^{J+1}(X_i)}\\ &= \sum_{i=1}^{M}\log\frac{(1-\alpha)P^{J}(X_i,c_i)+\alpha f(X_i,c_i)}{(1-\alpha)P^{J}(X_i)+\alpha f(X_i)} \qquad (22)\\ &= \sum_{i=1}^{M}\log\frac{P^{J}(X_i,c_i)+\epsilon f(X_i,c_i)}{P^{J}(X_i)+\epsilon f(X_i)}\end{aligned}$$

where $P^J(X_i) = \sum_{c=1}^{C}P^J(X_i,c)$, $f(X_i) = \sum_{c=1}^{C}f(X_i,c)$, and $\epsilon = \frac{\alpha}{1-\alpha}$.

A practically feasible optimization method to determine both $f(\cdot,\cdot)$ and ϵ, is expanding $\log P^{J+1}(\cdot|\cdot)$ into a Taylor's series around $P^J(\cdot,\cdot)$ with ϵf as the deviation (or increment) from $P^J(\cdot,\cdot)$, and ignoring the higher order terms as

$$\begin{aligned}\sum_{i=1}^{M}\log P^{J+1}(c_i|X_i) &\doteq \sum_{i=1}^{M}\log P^{J}(c_i|X_i)\\ &\quad+\epsilon\sum_{i=1}^{M}\frac{\partial\log P^{J+1}(c_i|X_i)}{\partial P^{J}(X_i,c_i)}f(X_i,c_i)\\ \doteq \sum_{i=1}^{M}\log P^{J}(c_i|X_i) &+ \epsilon\sum_{i=1}^{M}\frac{1-P^{J}(c_i|X_i)}{P^{J}(X_i,c_i)}f(X_i,c_i)\\ &= \sum_{i=1}^{M}\log P^{J}(c_i|X_i)+\epsilon\sum_{i=1}^{M}h_i f(X_i,c_i).\end{aligned} \qquad (23)$$

The approximate identity

$$\frac{\partial\log P^{J+1}(c_i|X_i)}{\partial P^{J}(X_i,c_i)} \doteq \frac{1-P^{J}(c_i|X_i)}{P^{J}(X_i,c_i)} \triangleq h_i \qquad (24)$$

is derived in the Appendix. Note that samples with a small posterior probability will receive a larger weight, *i.e.*, the samples not well accounted for under the current model will be paid more attention through weight h. For this reason the h_i's can be regarded as 'discriminative weights'.

It is now clear that once $\sum_{i=1}^{M}h_i f(X_i,c_i)$ is maximized we can easily find the best ϵ (or α) such that the posteriori probability is maximized. Therefore, the key optimization is to maximize

$$\sum_{i=1}^{M}h_i f(X_i,c_i) = \sum_{i=1}^{M}h_i\pi_{c_i}p_{c_i}(X_i) \qquad (25)$$

by determining the corresponding $\mu_{c_i},\sigma_{c_i},z_{c_i}$ (*i.e.*, $\mu_c,\sigma_c,z_c,c=1,2,\cdots,C$), taking discriminative weights h_i into account.

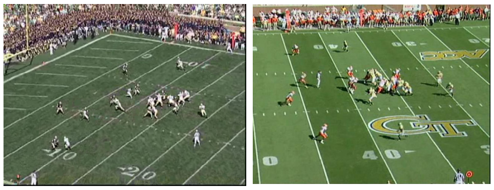

Figure 3. Samples of football plays used in the experiment.

This 'discriminatively weighted parameter estimation' for each component falls well into the EM framework. The E-step for this iterative procedure is $w_i = (h_i f(X_i,c_i))/(\sum_{i=1}^{M}h_i f(X_i,c_i))$ and the M-step is essentially to maximize $\sum_{i=1}^{M}w_i\log f(X_i,c_i)$ w.r.t. μ_c,σ_c,z_c. The optimal μ_c here, is exactly the weighted Karcher mean with weights w_i introduced in Section 3.2. Therefore, to implement the M-step we need and only need to perform the parameter estimation presented in 3.2 . After f is determined in this way, a line search for the best ϵ is followed to achieve the maximum $\sum_{i=1}^{M}\log P^{J+1}(c_i|X_i)$. If no ϵ can improve the discriminativeness, the algorithm terminates and the final number of components is J.

The multi-modal density learning presented above follows the line of recent work on boosting non-discriminative density functions [20] and is also inspired by the discriminative boosting for sequence classification [14]. However, in this work we are investigating a multi-class multi-modal probabilistic model for classification on a nonlinear manifold rather than in the Euclidean space, and in particular, applying the method to group activity recognition.

5. Experiment

The learning and recognition framework described above has been implemented on a collection of NCAA football games (sample snapshots in Figure 3). The play types have been annotated and the time span for each play is marked by an experienced person. The locations of each player are measured at time instants equally sampled within the time span. A constant amount of time duration is used for all plays so as to maintain a constant size of temporal interaction matrix. In particular, for activity samples with varing lengths, we always normalize their time scales to T (We set $T = 10$), with trajectory interpolation and temporal resampling if necessary. To find the time span of the occurrence of an activity automatically, *i.e.*, group activity detection, is not considered in this work.

We perform experiments on ground-truth data, where tracks for each player are manually labeled. The eleven trajectories for the eleven players on the offensive side are thus obtained for each play. Once a reliable multi-object tracker (with object role identification) is available, it can be incorporated for realizing a joint-tracking-and-recognition

Table 1. The confusion matrix of play recognition: H,C,M,L, and R stand for HITCH Dropback, Combo Dropback, Middle Run, Wideleft Run, and Wideright Run respectively.

	C	H	M	L	R
C	93.1	0.2	5.6	0.4	0.7
H	0.1	71.0	6.5	15.2	7.2
M	1.6	1.3	77.2	12.4	7.5
L	0.4	0.6	0.6	95.4	3.0
R	0.8	0.1	0.1	2.5	96.5

system. Though view points vary among different plays, no geometric transformation is applied since view-stability will be enforced when learning the optimal tensor reduction kernel. However, we do put the origin at the center of the objects and normalize the distances between the objects and the center. From more than a hundred play samples we select five play types, including *Combo Dropback, HITCH Dropback, Middle Run, Wideleft Run* and *Wideright Run*, totaling a number of 56 play samples. Other play types with too few samples are not considered. To get a sufficient amount of training samples, we generate multiple new play samples from different views for each of the existing plays. To achieve this we apply view transformations to each of the 56 samples, with 12 typical views selected from the original dataset. The view transformations are simply locally affine ones whose parameters are determined by locating the landmark points of the football field. Learning and then classification runs a multiple of times independently, each of which uses a random division of sample collection into training (80%) and testing (20%) sets. Other free parameters (*e.g.*, α, β in Section 2) in the framework are determined by experimental evaluation.

The average confusion matrix is shown in Table 1, indicating the percentage by which a specific play type is recognized as itself/another. An average recognition rate of 87.9% is observed from the confusion matrix. The fully quantitative comparison with previous work, especially [10], is difficult due to a completely different framework as well as different datasets being used. Note that the previous work is based on Bayesian network modeling with explicit domain knowledge about football game being incorporated. In contrast, the model in this paper works in a data-driven manner and thus easily extendable to other general coordinated group activities.

To evaluate the effectiveness of optimal tensor reduction as well as probabilistic modeling on DTIM, a comparative study is carried out with a baseline descriptor and three baseline classifiers. The baseline descriptor to compare with is the 'nominal' temporal interaction matrix obtained from the trivial tensor reduction kernel - the identity matrix. The baseline classifiers are selected as two nearest neighbor (NN) classifiers and supporter vector machine (SVM) classifier. One of the two NN classifiers defines the distance between two temporal interaction matrices as the usual Euclidean distance (NN Euclidean). The other, instead, makes use of the intrinsic distance on DTIM (NN on DTIM). The SVM classifier is employed from libSVM [1] where the multi-class classifier is implemented as a set of one-to-one binary ones. In each of these SVMs a radial basis function kernel is used together with the default parameter settings of the software. Classification is performed by taking majority of the votes from individual SVMs.

Table 2. Comparison of recognition performance (%).

	baseline	optimized
NN Euclidean	73.2	83.7
NN on DTIM	75.8	84.5
SVM	69.3	79.8
Probabilistic modeling on DTIM	76.3	87.9

The overall correct recognition rate is shown in Table 2. In all cases the improvement brought by optimized tensor reduction is clear. On the other side, by comparing probabilistic modeling on DTIM with the other three baseline classifiers, we actually investigated its advantage over three typical philosophies besides a Bayesian network paradigm. The NN Euclidean classifier ignores the intrinsic geometry of DTIM but regards all temporal interaction matrices as elements in Euclidean spaces. The SVM takes into account the probable nonlinear phenomenon in the Euclidean space but bypasses it with the kernel trick to pursue linear seperability. NN on DTIM, meanwhile, exploits the essential geometry of the data space without a statistical point of view. The comparison among the four demonstrates an empirical performance merit of the combination of both geometrical modeling and statistical modeling. Note that NN is only slightly weaker than proposed framework due to the relative 'flatness' of DTIM. Geometrical and probabilistic modeling on more 'curved' manifold will potentially achieve more significant performance gain.

6. Conclusion

In this work we investigated the modeling and recognition of coordinated multi-object activity in a data-driven manner. In particular, we proposed a temporal interaction matrix to characterize a group activity view-stably and discriminatively. We established the Riemannian geometry for the space of temporal interaction matrices, DTIM, and set up the 'intrinsic' probabilistic mechanism for random samples on DTIM. To better approximate the possibly complex distribution on DTIM, we further recursively built multi-component densities on DTIM in a way that inter-class separability is enhanced. We demonstrated the effectiveness of the proposed framework using football plays as experimental data. We made little use of football-specific domain knowledge and the framework is more generally extensible to other types of group activities.

Beyond this initial attempt to address multi-object activity problem, in the future we will investigate the following. The temporal interaction matrix relies on correctly obtaining point trajectories, which brings the issue of a more robust descriptor from incomplete/errorneous trajectories. A strict view-invariant approach, beyond an empirically view-stable descriptor remains a challenge. As has been mentioned, detection and segmentation of a particular group activity pattern, especially with a changing number of involved objects, is also of interest. Moreover, it is also interesting to look into activities from other domains other than football games. Last but not least, the study on probabilistic modeling on a nonlinear manifold is still in its infancy and thus worth further study.

7. Appendix: Derivation of (24)

By elementary calculus it is obvious that

$$\frac{\partial \log P^{J+1}(c_i|X_i)}{\partial P^J(X_i, c_i)} = \frac{\sum_{c=1,\cdots,C,c\neq c_i} P^J(X_i, c) + \epsilon \sum_{c=1,\cdots,C,c\neq c_i} f(X_i, c)}{(P^J(X_i, c_i) + \epsilon f(X_i, c_i))(P^J(X_i) + \epsilon f(X_i))}. \tag{26}$$

Since ϵf is a local deviation from P^J in Taylor's expansion, here we may assume $\epsilon f \ll P^J$. Hence we ignore the terms of ϵf and have the approximation

$$\begin{aligned}\frac{\partial \log P^{J+1}(c_i|X_i)}{\partial P^J(X_i, c_i)} &\doteq \frac{\sum_{c=1,\cdots,C,c\neq c_i} P^J(X_i, c)}{P^J(X_i, c_i) P^J(X_i)} \\ &= \frac{\sum_{c=1,\cdots,C,c\neq c_i} P^J(c|X_i)}{P^J(X_i, c_i)} = \frac{1 - P^J(c_i|X_i)}{P^J(X_i, c_i)}.\end{aligned} \tag{27}$$

Note that the best ϵ is determined after f is learned.

Acknowledgement: Videos are by courtesy of Georgia Tech Athletic Association. Ground-truth data are extracted and annotated by School of Interactive Computing, GA Tech. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-08-C-0047.

References

[1] *http://www.csie.ntu.edu.tw/ cjlin/libsvm/*. 7

[2] J. Aggarwal and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999. 1

[3] S. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000. 4

[4] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:781 – 796, 2000. 2

[5] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. John Wiley and Sons, 1998. 2

[6] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, 2003. 1

[7] A. Hakeem and M. Shah. Learning, detection and representation of multi-agent events in videos. *Artificial Intelligence*, 171:586 – 605, 2007. 1

[8] S. Hongeng and R. Nevatia. Multi-agent event recognition. In *ICCV*, 2001. 1

[9] C. Huang, H. Shih, and C. Chao. Semantic analysis of soccer video using dynamic bayesian network. *IEEE Transactions on Multimedia*, 8(4):749 – 760, 2006. 2

[10] S. Intille and A. Bobick. Recognizing planned, multiperson action. *Computer Vision and Image Understanding*, 81:414 – 445, 2001. 1, 2, 4, 7

[11] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez. Cross-view action recognition from temporal self-similarities. In *ECCV*, 2008. 2, 3

[12] R. Kass and P. Vos. *Geometric Foundations of Asymptotic Inference*. John Wiley and Sons, 1997. 4

[13] S. M. Khan and M. Shah. Detecting group activities using rigidity of formation. In *ACM Multimedia 2005*, 2005. 1

[14] M. Kim and V. Pavlovic. Discriminative learning of mixture of bayesian network classifiers for sequence classification. In *CVPR*, 2006. 6

[15] M. Lazarescu and S. Venkatesh. Using camera motion to identify different types of american football plays. In *ICME*, pages 181 – 184, 2003. 2

[16] T. Liu, W. Ma, and H. Zhang. Effective feature extraction for play detection in american football video. In *MMM*, 2005. 2

[17] X. Liu and C. Chua. Multi-agent activity recognition using observation decomposedhidden markov models. *Image and Vision Computing*, 24(2):166 – 175, 2006. 1

[18] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104:90–126, 2006. 1

[19] X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127 – 154, 2006. 5

[20] S. Rosset and E. Segal. Boosting density estimation. In *NIPS*, 2002. 6

[21] A. Srivastava, I. Jermyn, and S. Joshi. Riemannian analysis of probability density functions with applications in vision. In *CVPR*, 2007. 4

[22] N. Vaswani, A. Roy-Chowdhury, and R. Chellappa. Shape activity: A continuous-state hmm for moving/deforming shapes with application to abnormal activity detection. *IEEE Transactions on Image Processing*, 14:1603 – 1616, 2005. 1

[23] A. Veeraraghavan, R. Chellappa, and M. Srinivasan. Shape and behavior encoded tracking of bee dances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):463 – 476, 2008. 1

[24] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):1–45, 2006. 1, 2

[25] Y. Zhou, S. Yan, and T. S. Huang. Pair-activity classification by bi-trajectories analysis. In *CVPR*, 2008. 2