RECOGNIZING COORDINATED MULTI-OBJECT ACTIVITIES USING A DYNAMIC EVENT ENSEMBLE MODEL

Ruonan Li and Rama Chellappa

Center for Automation Research, University of Maryland, College Park, MD 20742, USA

ABSTRACT

While video-based activity analysis and recognition has received broad attention, existing body of work mostly deals with single object/person case. Modeling involving multiple objects and recognition of coordinated group activities, present in a variety of applications such as surveillance, sports, biological records, and so on, is the main focus of this paper. Unlike earlier attempts which model the complex spatial temporal constraints among different activities of multiple objects with a parametric Bayesian network, we propose a dynamic 'event ensemble' framework as a data-driven strategy to characterize the group motion pattern without employing any specific domain knowledge. In particular, we exploit the Riemannian geometric property of the set of ensemble description functions and develop a compact representation for group activities on the ensemble manifold. An appropriate classifier on the manifold is then designed for recognizing new activities. Experiments on football play recognition demonstrate the effectiveness of the framework.

Index Terms— Video Analysis, Activity Recognition

1. INTRODUCTION

This work deals with modeling and recognition of *coordinated group activities* involving *multiple objects* from videos. Human activity analysis and classification has been studied for nearly two decades [1, 2], focusing on single object cases. Activities of multiple objects exist widely in surveillance applications, sports, biological observation records, and consequently modeling and analysis of multi-object activities will be of use in these applications. In a less complex scenario, the individuals undergo structurally fixed motion[3] or follow similar dynamics or trajectories[4]. However, more meaningful and interesting semantics may be extracted for coordinated activities involving multiple objects. In other words, the individual objects will have distinctive and varying motion patterns but the group collectively demonstrates an underlying activity with an explicit semantic identity. A most illustrative example is a football game, in which we would like to recognize the strategy used rather than players' individual movements.

A group activity usually occurs according to a planned goal. The action of each object, meanwhile, is also the result of interactions with and response to the motion of other objects. The collaboration and interaction visually appear as a *temporally constrained co-occurrences* of individual motion primitives. In other words, a group activity is a collection of single-object activities occurring simultaneously or in a particular temporal order. Modeling and recognition of the temporal relationship (i.e. the group activity pattern) has been mostly handled using a Bayesian net framework[5, 6, 7, 8, 9]. Bayesian formulation, though successfully applied to modeling activities of a single object or motion, has some drawbacks while dealing with multi-object 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 works (e.g. [6]), where individual objects' ID's, roles and their individual action primitives were pre-labeled. Indeed, simultaneous recognition of individual actions and group activity pattern is intensive and prohibitive in some sense. Compared to the size of the state space and feature space of the network, the training data is insufficient most of the time. Thus not only the probabilistic dependence might very possibly be 'over-fitted', but also necessary priors are hard to learn from available data.

The work most similar to ours [6] designed large connected Bayesian networks for football play recognition. In contrast, we are exploring a 'data-driven' approach. Specifically, we regard a multi-object activity as a dynamically evolving ensemble of events. By an event we mean a trajectory or motion segment of a single object. Instead of identifying objects, we treat the events non-discriminatively as an event ensemble. However, we do learn an event vocabulary, which classifies events into different 'words'. Then for each time instant, we construct a word-space co-occurrence function, which characterizes the spatial distribution of different event words. The most desirable feature is that once given a proper Riemannian metric, the set of co-occurrence functions becomes exactly a Riemannian manifold. A group activity, i.e., a dynamic event ensemble, consequently becomes an evolution process on this manifold. The temporal relationship, which determines the activity pattern, is described by the evolving path of the co-occurrence function on the manifold. Eventually, modeling and recognition of coordinated group activity is achieved by statistical learning and inference on the manifold. The approach does not assume or pursue any causal structure and dependence of features and states of various levels.

2. LEARNING AN EVENT VOCABULARY

Although in some cases identifying the role of an object (e.g. the quarterback in a football play) is of help in predicting the group

intention, a non-discriminative scenario is more common for more general group activities, and thus more flexible and extendable. On the other hand, the event, i.e. the motion of an individual object can also be interpreted at different levels. Here we take the point trajectories, although body part movement or appearance can also be incorporated to provide a more powerful representation.

In this work, we assume that trajectories of individual objects are available, and denote a particular one as $X_t, t \in \mathfrak{T} = \{t_0, t_1, \cdots, t_f\}$. We do not pre-label the trajectories, but will cluster them in an unsupervised manner. To obtain trajectories we may need to employ a multi-object tracking module, but in this work we assume that such tracks are already obtained.

The collection of trajectory segments is to be evaluated in a pairwise manner to define a (dis)similarity index between every pair. The trajectories collected may take place with different starting times and locations, but the shape of the curves are actually the same, representing the same event primitive. Therefore, a temporal and spatial alignment of the trajectory is necessary. With one trajectory X_t defined above and another trajectory defined as $Y_s, s \in \mathfrak{S} = \{s_0, s_1, \cdots, s_g\}$ we define the aligned trajectories with respect to Y as

$$X'_{t,T} = X_{t+t_0+T} - X_{t_0+T}, t \in \mathfrak{T}' = \{0, 1, 2, \cdots, t_f - t_0 - T\}$$

where $T = 0, 1, 2, \cdots, T_{max}$ and

$$Y'_s = Y_{s+s_0} - Y_{s_0}, s \in \mathfrak{S}' = \{0, 1, 2, \cdots, s_g - s_0\}$$

By this temporal and spatial shifting, we pick up a subsegment of X and relocate it as well as Y at the S-T origin. In the same way the aligned trajectories X''_t and $Y''_{s,S}$ are also obtained.

The spatial-temporal alignment enables us to shift the primitive pair by controlling T and S for the best match. Under a fixed T, a straightforward dissimilarity measure between X and Y can be

$$\tilde{d}_{X \mapsto Y}(T) = \frac{1}{|\mathfrak{T}' \cap \mathfrak{S}'|} \sum_{r \in \mathfrak{T}' \cap \mathfrak{S}'} ||X'_{r,T} - Y'_r||$$

where $|.|$ denotes cardinality. By dividing the total distance or X' and Y' by the cardinality of $\mathfrak{T}' \cap \mathfrak{S}'$, we get an average dissimilarity measure between X and Y. However, by only considering the points in $\mathfrak{T}' \cap \mathfrak{S}'$ we also achieve robustness toward fragmented trajectories. Practically we may be unable to locate each object at every instant. By ignoring the time instants not shared by both trajectories, we are able to keep a lower dissimilarity between two similar but fragmented primitives. Similarly, under a fixed S, dissimilarity of Y and X is obtainable as $\tilde{d}_{Y \mapsto X}(S)$.

A by-product of ignoring the unshared time instants is that we may achieve a lower dissimilarity for a trajectory pair with less temporal overlap, though they may be actually very dissimilar. In fact, as the amount of overlap between the two increases, the confidence we have about the dissimilarity measure increases. Therefore, it is necessary to weight the above measure with a normalizing factor, resulting in the following confidence-weighted measure

$$\hat{d}_{X \mapsto Y}(T) = \tilde{d}_{X \mapsto Y}(T) \exp(1 - \frac{|\mathfrak{T}' \cap \mathfrak{S}'|}{|\mathfrak{T}|})$$

and the corresponding $\hat{d}_{Y \mapsto X}(S)$. It can be seen that as the overlap between the two trajectories increases, the weight is reduced, and consequently we are more confident about the dissimilarity measure.

Finally, to look for the best match between any primitive pair with varied S and T, the ultimate pairwise dissimilarity measure is taken as

$$d(X, Y) = \min\{\min_T \hat{d}_{X \mapsto Y}(T), \min_S \hat{d}_{Y \mapsto X}(S)\}$$

which will be fed to the unsupervised clustering procedure.

We only take into account the path distance in the above treatment. As mentioned before, the first and second order derivatives of X and Y and so on may be strong features to describe the event primitives, and accordingly the above distance can be easily extended to include more features. Another possible approach is to employ a time warping when comparing two trajectories as reported in [10]. Here we assume that an event is not a static but a *time-indexed* piece of trajectory, meaning that two events are different when they are executed with varied rate. Consequently time warping is not performed. An acceleration and a deceleration, for example, though along the same route within the same time, are not regarded as the same in our treatment.

To obtain a vocabulary for the event primitives, we recursively cluster all training trajectories into subsets, each of which is identified as an event word and those in the subset are treated as instances of the same word. Starting from pairwise dissimilarity, we employ the multiple-pass quadratic programming strategy in [11]. Its advantages over spectral clustering (e.g. Ncut [12]) are that it does not suffer from unstable eigenvector problem; and can automatically determine the best vocabulary size. In addition, a simple optimization algorithm is available for it.

3. DYNAMIC EVENT ENSEMBLE AS AN EVOLUTION PROCESS ON MANIFOLD

With a learned vocabulary of all possible event primitives, a coordinated group activity is essentially an ensemble of event 'words' dynamically evolving with time. Meanwhile, individual events may occur at different spatial locations, and different spatial configurations and distribution of event words will imply different activity types. Therefore, a group activity is completely characterized by a time series of event word distribution.

The *spatial co-occurrence function* $f(w, \Omega; t)$ at time t is defined as the occurrence intensity of event $w \in \{1, 2, \cdots, W\}$ in a spatial kernel $\Omega \in \prod \subseteq (\Omega_p, \Omega_q)$, where $\prod$ is the set of indices of all spatial kernels which cover the entire spatial range of possible occurrence. Occurrence intensity may be interpreted in different ways, and here we regard the total counts of the occurrences of an event w in the kernel Ω as the occurrence intensity. A co-occurrence function corresponds to a possible spatial event distribution. However, for a fixed group size, the set of all co-occurrence functions is not an Euclidean space. One may want to expand the set to account for the varying group size. Nevertheless, this is contradictory to the notion that a change in group size normally indicates an activity boundary, meaning, an addition or deletion of an object terminates the ongoing

activity and initialize a new one. Thus, we address the coherent activity pattern rather than change detection, and limit our attention to activities involving a fixed number of objects.

The non-Euclidean property of set of co-occurrence functions must be accounted for in mathematical formulation and learning. We first apply a normalization step to get a normalized co-occurrence function

$$G(w,\Omega;t) = \frac{(f(w,\Omega;t))^{\frac{1}{2}}}{(\sum_{w=1}^{W}\int_{\Omega_p}^{\Omega_q} f(w,\Omega;t)d\Omega)^{\frac{1}{2}}}$$

and the set of normalized co-occurrence function becomes a Riemannian manifold. For any two elements g_1, g_2 in the tangent space $\mathscr{T}_G$ at G, the Riemannian metric is defined as

$$< g_1, g_2 > \triangleq \sum_{w=1}^{W}\int_{\Omega_p}^{\Omega_q} g_1(w,\Omega;t)g_2(w,\Omega;t)d\Omega$$

A related but simpler case was detailed recently in [13]. For clarity we call this manifold the *event ensemble manifold*. With the above defined Riemannian metric, the basic geometry of the event ensemble manifold is straightforward (we omit these discussions due to page limitation), and serves as powerful tools for learning and inference.

A single spatial co-occurrence function is a 'holistic' but static description of event ensemble, not taking temporal evolution or time constraint into account. However, the coordinated group activity is essentially a dynamic ensemble of events, and it is this temporal process that critically determines the specific semantic pattern. Therefore, time-series modeling is a natural and necessary step towards group activity characterization. The time sequence of co-occurrence functions, nevertheless, is not a sequence in Euclidean space, where we have rich and powerful tools on hand.

Let us denote $G(w,\Omega;t)$ by $G(t)$ from now on for simplicity, and keep in mind that as t varies $G(t)$ is an evolving process. However, beyond this we can hardly make any stronger assumptions. In this case, to find a proper quantitative feature for each group activity pattern we are going to use the *activity characteristic curve* defined as

$$C(t) = \mathbb{E}(G(t))$$

which is nothing but the mean value curve for the evolution process $G(t)$. With multiple training sequences for the same type of group activity, we are always able to find the mean sequence without additional assumptions, though the mean should be obtained from an average on the manifold rather than in Euclidean space. It can be expected that different activity characteristic curve corresponding to different activities will be located distinctively on the event ensemble manifold, therefore providing us the capability to classify a new sequence.

The activity characteristic curve $C(t)$, i.e. the expectation curve on the manifold is explicitly defined as

$$C(t) = \arg\min_{G} \mathbb{E}(d^2(G, G(t)))$$

according to [14], where d is the intrinsic distance on the manifold induced by the Riemannian metric rather than the usual Euclidean distance $||G - G(t)||$. If there exists a probability density function $p(G(t))$ for $G(t)$, then we have

$$C(t) = \arg\min_{G} \int_{\mathscr{M}} d^2(G, G(t))p(G(t))d\mathscr{M}$$

where $d\mathscr{M}(t)$ can be thought as a 'patch' of the manifold. Note that the integration is performed on the manifold only. Since we do not make more assumptions about p, practically we estimate the density function by kernel method using a set of training samples $\{(G_i(t), t_i)\}_i$

$$p(G(t)) = \frac{\sum_i \mathbb{K}(d^2(G(t), G_i), t - t_i)}{k(t)}$$

For simplicity we make use of the Nadaraya-Watson kernel [15][16]so that $\mathbb{K}$ can be separated into a temporal factor and a spatial kernel

$$p(G(t)) = \frac{\sum_i \mathbb{K}_{H_s}(d^2(G(t), G_i))\mathbb{K}_{H_t}(t - t_i)}{k(t)}$$

where H_s and H_t are the spatial and temporal kernel bandwidths respectively.

To perform the above minimization we adopt the general theory presented in [14] to our specific event ensemble manifold. By finding the mean co-occurrence function $C(t)$ at all t, we eventually obtain the activity characteristic curve for each group activity.

4. NACC CLASSIFIER FOR NEW ACTIVITY

With c types of group activities represented as $\{C_i(t)\}_{i=1,2,\cdots,c}$, we are in a position to categorize a new incoming activity into one of these classes. The classification is straightforwardly achieved in two steps. In the first step, we should identify each of the event primitives in the incoming video as one of the event words. Then we construct the evolving event co-occurrence function sequence $D(t)$ and classify it into one of the activity types. For the former, we label a new event as the word with which it shares the most similarity defined in Section 2. For the latter we use the Nearest Activity Characteristic Curve (NACC) classifier as

$$Activity(D(t)) = \arg\min_{j} \sum_{t} d(D(t), C_j(t))$$

The classifier looks for the activity type with minimum manifold distance from the one in the testing video clip. Also, the classifier can be interpreted as a correlator, which picks up the maximum manifold correlation as the recognition result. In the language of signal processing, this can be viewed as matched filtering on a manifold.

5. EXPERIMENT

The learning and recognition framework described above has been implemented on a collection of NCAA football games. In the videos of the games, the play types have been annotated, the time span for each play is marked by an experienced football player, and tracks for each player are also marked. Apparently, once a reliable multi-object tracking module is available, it can be incorporated. Here we

Table 1. Confusion matrix of play recognition on real data: H,P,R,S, and T stand for HITCH, POWO, REDHAG, STRCH, and TSMIKE respectively.(%)

	H	P	R	S	T
H	72.2	7.4	13.7	0.0	6.7
P	3.6	64.3	0.7	22.9	8.5
R	2.5	1.1	92.8	1.6	2.0
S	3.3	5.3	0.9	82.2	8.3
T	3.7	0.0	9.0	0.0	87.3

use manual labeling to focus on activity pattern analysis. Even so, incomplete trajectories occur commonly when players are occluded or move beyond camera range. To account for zooming and panning effects and get trajectories in field coordinates, a geometric transformation determined by locating the field landmarks is applied to all points in each frame.

From more than hundreds of play samples we select five play types, including *HITCH, POWO, REDHAG, STRCH* and *TSMIKE*, each of which contains enough samples. Learning and then classification algorithms are run multiple times, each run using a random division of sample collection into training and testing sets. For the same play with different formations, a separate activity characteristic curve is learned for each formation. The missing trajectory effect is exactly handled with the proposed pairwise similarity measure. 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 correct recognition rate of 80% is observed from the confusion matrix. The fully quantitative comparison with previous work, especially [6], is difficult due to a completely different framework, unavailability of implementation details, as well as different dataset being used. However, qualitatively it is seen from Figure 13 in [6] that we achieve a recognition performance no worse than [6] . Note that the previous work uses parametric Bayesian network modeling with explicit domain knowledge about football game incorporated. In contrast, the event ensemble model in this paper works autonomously and is directly extendable to other coordinated group activities.

The main computational load comes from generating the pairwise similarity matrix together with new event categorization. The vocabulary discovery and activity characteristic curve learning converge quickly in several iterations.

6. CONCLUSIONS

In this work we recognize a coordinated multi-object activity using a dynamic event ensemble framework. We first iteratively learn a vocabulary for single-object motion (event) patterns with pairwise relationship tailored to account for non-robust feature extraction. Then naturally from the Riemannian property of the set of all co-occurrence functions (spatial event distributions) we look into this event ensemble manifold consisting of event words, and find a compact representative subset (curve) for each group activity type. By eventually representing group activities as the characteristic curves, we exploit the metric on the manifold and develop manifolded-nearest-neighbor classifier to recognize new activities.

7. REFERENCES

[1] J.K. Aggarwal et. al., "Human motion analysis: a review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[2] T. B. Moeslund et. al., "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, pp. 90–126, 2006.

[3] S. M. Khan et. al., "Detecting group activities using rigidity of formation," in *ACM Multimedia*, 2005.

[4] N. Vaswani et. al., "Shape activity: A continuous-state hmm for moving/deforming shapes with application to abnormal activity detection," *IEEE Trans. Image Processing*, vol. 14, pp. 1603 – 1616, 2005.

[5] S. Hongeng et. al., "Multi-agent event recognition," in *ICCV*, 2001.

[6] S.S. Intille et. al., "Recognizing planned, multiperson action," *Computer Vision and Image Understanding*, vol. 81, pp. 414 – 445, 2001.

[7] S. Gong et. al., "Recognition of group activities using dynamic probabilistic networks," in *ICCV*, 2003.

[8] X. Liu et. al., "Multi-agent activity recognition using observation decomposedhidden markov models," *Image and Vision Computing*, vol. 24, no. 2, pp. 166 – 175, 2006.

[9] A. Hakeem et. al., "Learning, detection and representation of multi-agent events in videos," *Artificial Intelligence*, vol. 171, pp. 586 – 605, 2007.

[10] X. Wang et. al., "Learning semantic scene models by trajectory analysis," in *ECCV*, 2006.

[11] M. Pavan et. al., "Dominant sets and pairwise clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 167 – 172, 2007.

[12] J. Shi et. al., "Normalized cuts and image segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888 – 905, 2000.

[13] A. Srivastava et. al., "Riemannian analysis of probability density functions with applications in vision," in *CVPR*, 2007.

[14] X. Pennec, "Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements," *Journal of Mathematical Imaging and Vision*, vol. 25, no. 1, pp. 127 – 154, 2006.

[15] E. A. Nadaraya, "On estimating regression.," *Theory of Probability and its Applications*, vol. 25, pp. 186 – 190, 1964.

[16] G. S.Watson, "Smooth regression analysis," *Sankhya*, vol. 26, pp. 101 – 116, 1964.