Spatio-Temporal Learning of Basketball Offensive Strategies

Ching-Hang Chen
Inst. of Information Science
Academia Sinica
Nankang, Taipei 115, Taiwan

Tyng-Luh Liu
Inst. of Information Science
Academia Sinica
Nankang, Taipei 115, Taiwan

Yu-Shuen Wang
Dept. of Computer Science
National Chiao Tung University
Hsinchu 300, Taiwan

Hung-Kuo Chu
Dept. of Computer Science
National Tsing Hua University
Hsinchu 300, Taiwan

Nick C. Tang
Inst. of Information Science
Academia Sinica
Nankang, Taipei 115, Taiwan

Hong-Yuan Mark Liao
Inst. of Information Science
Academia Sinica
Nankang, Taipei 115, Taiwan

ABSTRACT

Video-based group behavior analysis is drawing attention to its rich applications in sports, military, surveillance and biological observations. The recent advances in tracking techniques, based on either computer vision methodology or hardware sensors, further provide the opportunity of better solving this challenging task. Focusing specifically on the analysis of basketball offensive strategies, we introduce a systematic approach to establishing unsupervised modeling of group behaviors. In view that a possible group behavior (offensive strategy) could be of different duration and represented by dynamic player trajectories, the crux of our method is to automatically divide training data into meaningful clusters and learn their respective spatio-temporal model, which is established upon Gaussian mixture regression to account for intra-class spatio-temporal variations. The resulting strategy representation turns out to be flexible that can be used to not only establish the discriminant functions but also improve learning the models. We demonstrate the usefulness of our approach by exploring its effectiveness in analyzing a set of given basketball video clips.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*video analysis*

Keywords

Group action recognition, machine learning

1. INTRODUCTION

Action/behavior analysis based on video content has been actively studied for decades. While quite a number of practical techniques are established mostly for dealing with single or two objects of concern, we focus especially on group behaviors involving multiple objects. Observe that the nature of this problem engages high-level semantics and reveals crucial information on the plan or goal of the associated group. For example, in the popular basketball sport, the coordination of players' movements on the court presents a well-thought-out pattern of group behavior, which is often regarded as a strategy. Analysis on these strategies helps the teams review the performance of their offense and defense executions, understand competitors' strategies, and highlight the weakness for future improvement reference. In addition, the discovered information would bring new insights into the game such as providing in-depth investigation on winning rate or evaluating value of certain players. However, the analysis typically requires onerous human labeling efforts, and expert knowledge on recognizing underlying strategies and interpreting the contents. Thanks to the advance in tracking techniques (vision-based or sensor-based), the activity of each individual can be extracted according to the intended features. Yet, even with the available data, group behavior analysis poses challenging tasks:

- The representation of a group behavior generally includes actions of multiple independent objects in terms of, *e.g.*, player trajectories. This could cause ambiguity in that their ordering might be randomly assigned, and needs to be addressed when evaluating the similarity with other data.
- To establish a robust description for a representative group behavior, we should find a way to account for the spatio-temporal variations of each involved individual among intra-class samples.
- Finally, to recognize the underlying group behaviors in a test video clip based on the learned representation models, designing the associated discriminant functions for classification is also a critical issue.

Existing relevant techniques for group behavior recognition are mostly model-based, and adopt, say, SVMs for classifications. Siddiquie *et al.* [11] consider *multiple kernel learning* with SVMs on football tactic classification. Ballan *et al.* [1] use SVMs with spatio-temporal features, generated from a *bag-of-words* model. In [15], Wang *et al.* instead employ SVMs with the designed motion descriptors. In addition to these efforts, probability-based approaches are also

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806297.

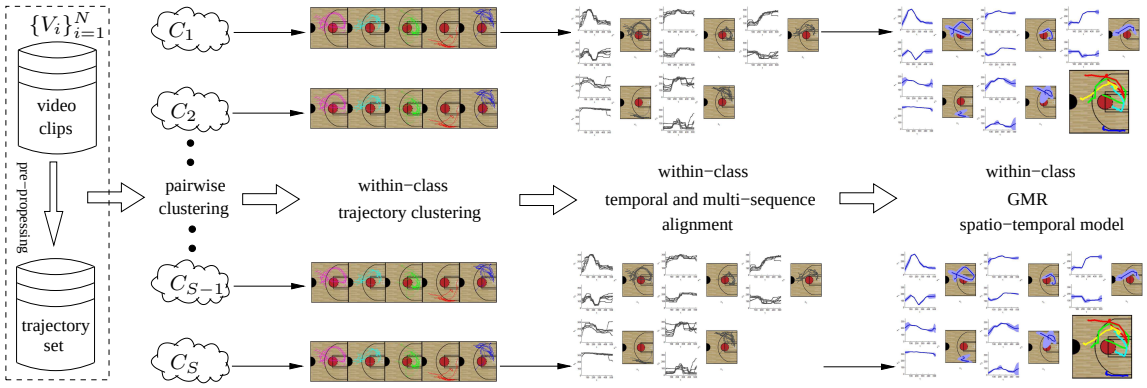


Figure 1: Flowchart of our method for learning the spatio-temporal models from given video clips.

popular. Li *et al*. [9] investigate *discriminative temporal interaction manifold* for football strategy recognition. Hu *et al*. [6] propose a *Dirichlet process mixture model* for trajectory analysis. Moreover, the Markov model is utilized in [13, 2, 12] for multi-trajectory behavior classification. The above-mentioned methods are implemented generally with supervised scenarios and thus require massive labeling efforts, while in our approach, models of group behavior can be learned in an unsupervised manner.

Figure 1 illustrates the steps of our method. We first represent each video clip in multi-trajectory form and measure pairwise trajectory-set distances by *dynamic time warping* (DTW) [7, 14, 16]. Leveraging with the pairwise clustering technique [10], we divide the data into clusters, each of which corresponds to a particular class of offensive strategy. On establishing the spatio-temporal models, we further partition each (strategy) cluster into (trajectory) sub-clusters, and perform temporal alignment as well as *Gaussian mixture regression* (GMR) [5] to complete the model learning. To accomplish classification, the learned GMR parameters can be incorporated into the resulting discriminant functions.

2. SPATIO-TEMPORAL LEARNING

We begin by proposing an efficient procedure to automatically divide a given collection of training video data into meaningful clusters. Denote the set of N video clips to be analyzed as $D = \{V_1, \ldots, V_N\}$ and assume they demonstrate totally S different offensive strategies. In case that the value of S is not available in training, our method can effectively overcome the difficulty by employing an iterative clustering scheme to yield a reasonable estimate. The goal of data clustering is then to divide D into S clusters so that our approach can establish their respective spatio-temporal model and discriminant function for classifying a test video clip.

2.1 Distance Measures

To accomplish the desired clustering, we first construct the representation of a basketball video clip and the distance measure. As each video records players' (half-court) actions during an interval of a basketball match, we pre-process the underlying image sequence using a *tracking by detection* [3] technique to obtain the players' trajectories of the offensive team. Specifically, we represent a video clip V, say, of F frames by $V = \{T_p\}_{p=1}^P$ where $P = 5$ reflects the number of players in an offensive team, and $T_p = \{\mathbf{x}_{p,f}\}_{f=1}^F \in \mathbb{R}^{d\times F}$ is the trajectory specified by the d-dimensional coordinates of player p at each image frame. Note that the exact ordering among the player trajectories is not specified as the ambiguity will be resolved via trajectory matching.

The trajectory-based representation implies that constructing a distance measure to correlate two video clips naturally relies on how the trajectory similarity is evaluated. For two arbitrary (player) trajectories $T \in \mathbb{R}^{d\times F}$ and $T' \in \mathbb{R}^{d\times F'}$, we take account that the two trajectories may not be of the same length, *i.e*., $F \neq F'$, and exploit the DTW algorithm to define an appropriate distance function by

$$d_t(T, T') = \min_{W_T, W_{T'}} \frac{1}{\ell} \|TW_T - T'W_{T'}\|_F \qquad (1)$$

where $\|\cdot\|_F$ is the Frobenius norm and ℓ, $W_T \in \{0,1\}^{F\times\ell}$, $W_{T'} \in \{0,1\}^{F'\times\ell}$ are the number of matchings and the warping matrices of a plausible dynamic-programming solution path. With (1), we set the distance function between two video clips, say, $V = \{T_p\}_{p=1}^P$ and $V' = \{T'_p\}_{p=1}^P$ as

$$d_c(V, V') = \min_{\pi(p)} \sum\nolimits_{p=1}^{P} d_t(T_p, T'_{\pi(p)}) \qquad (2)$$

where π is a permutation of $\{1, \ldots, P\}$.

2.2 Iterative Pairwise Clustering

We are to divide $D = \{V_1, \ldots, V_N\}$ into S clusters, each of which corresponds to a specific offensive strategy of basketball. As we prefer accommodating the situation that the exact value of S may not be known in advance, we consider the *dominant-set clustering* by Pavan and Pelillo [10]. In particular, we use (2) to yield the affinity matrix A and solve the following optimization problem:

$$\max_{\mathbf{y}} \ \mathbf{y}^T A\mathbf{y} \ \text{ subject to } \mathbf{y} \geq \mathbf{0} \in \mathbb{R}^N \text{ and } \mathbf{e}^T\mathbf{y} = 1 \qquad (3)$$

where the (i,j)th entry of A is given by $\exp\{-d_c(V_i, V_j)/\sigma^2\}$ and $\sigma > 0$ is a scale parameter. Performing clustering with (3) is carried out iteratively. It starts by treating each V_i in D as a graph node and specifying a pre-determined positive threshold δ. Having solved (3), those nodes corresponding to positive components of the optimal $\mathbf{y}$ will be removed from the graph, and included in a new cluster if their value is larger than δ. With the updated graph, we repeat the

above procedure until the algorithm uncovers S clusters, denoted as $C_1, \ldots, C_S$. If S is unknown, the iterative process can be terminated when reaching an empty graph. One notable advantage of the aforementioned clustering technique is that it provides a convenient way to exclude noisy data from corrupting the clustering result.

2.3 Spatio-temporal Representations

Our task now is to derive a spatio-temporal representation for each (strategy) cluster C. Suppose that C comprises Q video clips from D and thus has $P \times Q$ player trajectories. To yield an intra-cluster representation, we note that all the pairwise distances among player trajectories have already been evaluated when forming the affinity matrix A in (3). It follows that we can readily obtain the affinity matrix for the $P \times Q$ trajectories and again use dominant-set clustering to divide them into P (trajectory) clusters, upon which the strategy representation is established.

For each of the P clusters, let the underlying set of trajectories be $\{T_q\}_{q=1}^{Q}$ to reflect that they are from Q different video clips. We then carry out the following two steps.

1. We couple *multi-sequence* DTW [16] with *derivative* DTW [8] so that spatio-temporal relationships over sequences of different lengths can be more effectively explored. That is, we consider solving

$$\min_{\{W_q\}} Q \sum\nolimits_{q=1}^{Q} \|\delta T_q W_q - \frac{1}{Q} \sum\nolimits_{q'=1}^{Q} \delta T_{q'} W_{q'}\|_F \quad (4)$$

   where δT_q denotes the trajectory derivative and $\{W_q\}_{q=1}^{Q}$ are the warping matrices.

2. Assume that optimizing (4) results in q^* matchings between any two trajectories. We then apply Gaussian mixture regression to generate the mean trajectory $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_{q^*}] \in \mathbb{R}^{d \times q^*}$ with the variance matrix $\boldsymbol{\sigma}^2 = [\sigma_1^2, \ldots, \sigma_{q^*}^2] \in \mathbb{R}^{d \times q^*}$.

Finally, for each (strategy) cluster C, we can derive its spatio-temporal representation, denoted as $\{(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)\}_{p=1}^{P}$.

2.4 Refinement and Discriminant Functions

Having established the spatio-temporal representations, we can define for each (strategy) cluster C its discriminant function f for classifying a new video clip, say, $V' = \{T'_p\}_{p=1}^{P}$ where trajectory $T'_p = \{\mathbf{x}'_{p,f}\}_{f=1}^{F'} \in \mathbb{R}^{d \times F'}$. Let the underlying spatio-temporal representation be $\{(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)\}_{p=1}^{P}$. Then the discriminant function is defined by

$$f(V') = \min_{\pi(p)} \sum\nolimits_{p=1}^{P} \tilde{d}_t(T'_{\pi(p)}, \boldsymbol{\mu}_p) \quad (5)$$

and

$$\tilde{d}_t(T'_{\pi(p)}, \boldsymbol{\mu}_p) = \min_{W_{T'} \in \{0,1\}^{F' \times q^*}} \|\frac{1}{\boldsymbol{\sigma}_p} \circ (T'_{\pi(p)} W_{T'} - \boldsymbol{\mu}_p)\|_F \quad (6)$$

where $\tilde{d}_t$ is a modified trajectory distance function weighted by taking Hadamard product with respect to the matrix $1/\boldsymbol{\sigma}_p$, whose (i,j)th entry is just the reciprocal of the (i,j)th entry of $\boldsymbol{\sigma}_p$, to account for the distribution assumption of a GMR representation. Having obtained all the discriminant functions $f_1, \ldots, f_S$ for clusters $C_1, \ldots, C_S$, respectively, we are ready to lay out the discriminant rule to decide which (strategy) cluster should V' be associated with:

$$V' \mapsto C_{s^*} \text{ if } s^* = \arg \min_{s \in \{1,\ldots,S\}} f_s(V'). \quad (7)$$

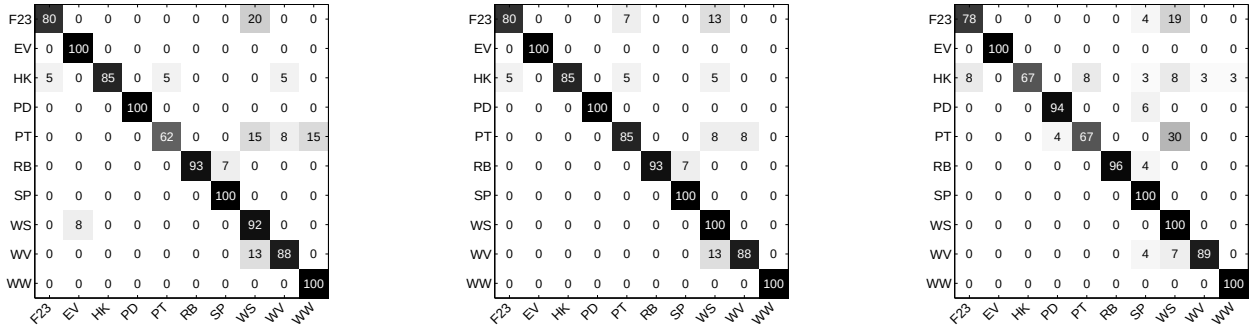
The discriminant rule in (7) is useful not just in testing but also in training. We can form an EM-like iterative training procedure to refine the spatio-temporal model. Specifically, we use the techniques just described to obtain the initial spatio-temporal representations and the discriminant functions. In the E-step, we use (7) to re-assign the training videos to their proper cluster. Then, in the M-step, we update each spatio-temporal model. We repeat the process until an M-step yields no re-clustering change.

So far, our proposed formulation is a fully unsupervised approach, but we indeed do not know what specific offensive strategies of basketball the S (strategy) clusters correspond to. To extend the usefulness of our method, we exploit *minimal* domain knowledge to label only S video clips, each of which is the one that has the least weighted distance, as in (6), to its cluster-wise mean trajectory. This way we could establish a classification framework for predicting the type of basketball offensive strategy without relying on manually labeling massive video clips.

3. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of our method, we carry out experiments on modeling and classifying basketball offensive strategies. Specifically, we collect totally 134 video clips from NBA games, including ten different types of basketball offensive strategies. They are termed as `2-3 Flex`, `Elevator`, `Hawk`, `Pin-down`, `Princeton`, `Back-side pick and roll`, `Side-pick slip and pop`, `Single`, `Weave` and `Wing-wheel`. For simplicity, they are abbreviated by F23, EV, HK, PD, PT, RB, SP, WS, WV and WW, respectively. Each video clip yields one associated trajectory set containing five player trajectories from the offensive team on the court, and depicts the execution of a particular strategy out of the ten offensive types. (The ground-truth labels are annotated by experts.) Literatures such as [4] also propose to automatically acquire intended trajectory data in basketball games, though this aspect is beyond the focus of our discussion.

The spatio-temporal models are learned via unsupervised learning. To verify the advantage of the EM-like iterative training procedure described in Section 2.4, we use the resulting discriminant functions to categorize the training data and generate the confusion matrix. In Figures 2(a) and 2(b), we show the confusion matrices of the training procedures with and without using the EM iteration. There we can see that the EM-like approach can achieve an average accuracy of about 93.1%, which significantly improves the non-EM training result. We then split the set of video clips that 80% of them are randomly selected for training and the remaining 20% for testing. The test video clips are classified according to (7), where the respective discriminant functions are learned by our method with the EM-like procedure. To achieve a more reliable evaluation, we repeat the testing ten times (with different data splits), and obtain an average classification rate of 89.1%. (See the average confusion matrix in Figure 2(c).) The good performance indicates that the proposed spatio-temporal learning is promising for uncovering the subtle group behaviors behind the complex dynamics of a very active sports event, such as an NBA basketball game.



(a) Without EM (b) With EM (c) Testing result

Figure 2: Confusion matrices from training and testing: F23,EV,HK,PD,PT,RB,SP,WS,WV,WW represent the 10 different basketball offensive strategies. (a) Learning without EM refinement. (b) Learning with EM refinement. (c) Average confusion matrix of strategy classification from 80% (training) vs. 20% (testing) random split of the dataset.

4. CONCLUSIONS

Learning to understand group behaviors from video data is challenging. Our method starts with a succinct trajectory-based representation to encode the highly-dynamic player actions, and establish a DTW-based distance function to measure the closeness between two video clips. These enable dividing the dataset into clusters so that we can separately learn their spatio-temporal model. To accommodate the intra-class variations, we explore the GMR technique to boost the robustness of the learned discriminant functions. Our experimental results are encouraging and it is promising to extend the proposed formulation to address analyzing group behaviors of other forms. For future research efforts, it would be interesting to generalize our method to handle that the group size could vary in different video clips.

5. ACKNOWLEDGMENTS

This work was supported in part by MOST grants 103-2221-E-001-009-MY3, 102-2221-E-001-021-MY3, 102-2221-E-007-055-MY3 and 103-2221-E-007-065-MY3.

6. REFERENCES

[1] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Video event classification using string kernels. *Multimedia Tools and Applications*, 48(1):69–87, 2010.

[2] F. I. Bashir, A. A. Khokhar, and D. Schonfeld. Object trajectory-based activity classification and recognition using hidden markov models. *Image Processing, IEEE Transactions on*, 16(7):1912–1919, 2007.

[3] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *IEEE International Conference on Computer Vision*, pages 1515–1522, 2009.

[4] H.-T. Chen, C.-L. Chou, T.-S. Fu, S.-Y. Lee, and B.-S. P. Lin. Recognizing tactic patterns in broadcast basketball video using player trajectory. *Journal of Visual Communication and Image Representation*, 23(6):932–947, 2012.

[5] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 1996.

[6] W. Hu, X. Li, G. Tian, S. Maybank, and Z. Zhang. An incremental DPMM-based method for trajectory clustering, modeling, and retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1051–1065, 2013.

[7] A. Kassidas, J. F. MacGregor, and P. A. Taylor. Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 44(4):864–875, 1998.

[8] E. J. Keogh and M. J. Pazzani. Derivative dynamic time warping. In *SDM*, volume 1, pages 5–7. SIAM, 2001.

[9] R. Li, R. Chellappa, and S. K. Zhou. Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2450–2457, 2009.

[10] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(1):167–172, 2007.

[11] B. Siddiquie, Y. Yacoob, and L. S. Davis. Recognizing plays in american football videos. Technical report, University of Maryland, 2009.

[12] G. Sukthankar and K. Sycara. Activity recognition for dynamic multi-agent teams. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(1):18, 2011.

[13] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2004–2011, 2009.

[14] M. Vlachos, G. Kollios, and D. Gunopulos. Discovering similar multidimensional trajectories. In *18th International Conference on Data Engineering*, pages 673–684. IEEE, 2002.

[15] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[16] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1282–1289, 2012.