Group Motion Segmentation Using a Spatio-Temporal Driving Force Model

Ruonan Li and Rama Chellappa
Department of Electrical and Computer Engineering, and Center for Automation Research, UMIACS
University of Maryland, College Park, MD 20742
{liruonan,rama}@umiacs.umd.edu

Abstract

We consider the 'group motion segmentation' problem and provide a solution for it. The group motion segmentation problem aims at analyzing motion trajectories of multiple objects in video and finding among them the ones involved in a 'group motion pattern'. This problem is motivated by and serves as the basis for the 'multi-object activity recognition' problem, which is currently an active research topic in event analysis and activity recognition. Specifically, we learn a Spatio-Temporal Driving Force Model to characterize a group motion pattern and design an approach for segmenting the group motion. We illustrate the approach using videos of American football plays, where we identify the offensive players, who follow an offensive motion pattern, from motions of all players in the field. Experiments using GaTech Football Play Dataset validate the effectiveness of the segmentation algorithm.

1. Introduction

In this work we consider the problem of *group motion segmentation*, and propose a solution for it. The group motion segmentation problem arises from video surveillance applications and sports video analysis, where a 'group motion' involving multiple participants is of interest. Specifically, we have in hand point trajectories from consecutive frames of a video sequence, and aim to group them into two or more clusters. While this may appear to be similar to the traditional motion segmentation problems [5, 7, 13, 16, 26, 33, 35, 36, 32], it is actually different in the following sense. During group motion, the participating objects/people have distinctive and varying motions but the group itself collectively demonstrates an underlying activity of a particular pattern, while the non-participating group of objects/people does not demonstrate that pattern. A recent development in the area of video analysis and activity recognition is the need for analyzing these motion patterns of the participating group, which are also called 'multi-object' or 'group' activities [17, 31, 12, 14, 21, 11, 8, 37, 23, 19, 24, 28], and various approaches have been proposed to recognize the group motion pattern or detect a change or an anomaly. However, these works assume that all objects are involved in the activity, which is far from realistic scenarios where only a portion of the objects/people contribute to the specific group motion. The group motion segmentation problem explored here attempts to divide the point trajectories into participating group and non-participating group, or into multiple participating groups, each corresponding to a group activity.

It is important to look into examples of 'group motion' patterns of interest and see where the challenges are. One can think of simple cases where the group of people parade toward the same direction with same speed [17], or passengers getting out of a plane and walking towards the terminal follow a stationary dynamics [31]. In these cases a deviation from the stationary model is detected as an 'event' or an 'anomaly'. In more complex problems, a limited number of participants interact with each other, and typical activities include approaching, chasing, hand-shaking, robbery, fighting, etc. [12, 21, 11, 37, 23, 24, 28]. Also, a group activity in an outdoor airport ramp [8] involves several individual activities occurring in a constrained area and in a specific temporal order. The most challenging case is American football plays involving a greater number of participants, where the task is to recognize the play strategy of the offensive players from their moving trajectories [14, 19]. In a football play, the offensive players are the participants of the group motion of offense while the defensive players are non-participants. Different offensive participants will give rise to different moving trajectories, while the group will collaboratively demonstrate an ensemble motion pattern which can be identified as a semantic strategy represented as a play diagram in the playbook. This group motion pattern manifests itself as the spatial constraint and temporal co-occurrence among the interacting trajectories. Note that participants move under the guidance of a play diagram but significant spatio-temporal variations exist among different realizations. Also, the participating and non-participating group are mixed within the same area, thus non-distinguishable by simply partitioning the spatial

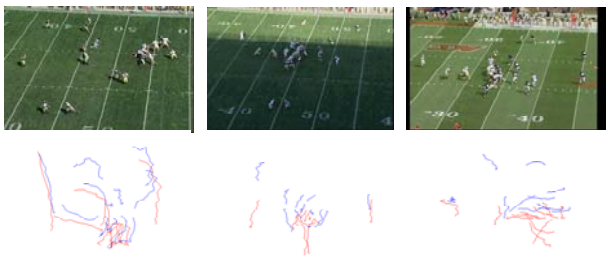


Figure 1. Samples from GaTech Football Play Dataset. The top row gives snapshots of videos of the plays and bottom row contains corresponding trajectories. The red trajectories are offensive ones and the blue defensive ones. The trajectories are displayed in ground plane coordinates for better visualization.

domain. For these reasons, we address the group motion segmentation problem in the context of offensive player identification in football plays. Examples of the group motion patterns under our consideration are given in Figure 1. Note that our segmentation is based on motion only: we do not make use of appearance based features, which may not be always available due to poor video quality or other reasons.

There are additional challenges beyond the aforementioned ones. Though the participating group of a football play consists of a constant number of objects, more generally the group motion pattern may be executed by a varying number of agents in different realizations, and the number may change during the activity due to a participant's departure or the arrival of a new member. Moreover, as the motion trajectories are generated by a tracking algorithm, they are noisy. The trajectories may be incomplete, fragmented or missing due to limitations of the tracking algorithm, strong occlusion among objects, and other issues such as video quality. Each of these challenges should be addressed by a vision algorithm and indeed our method is able to handle them.

Turning back to traditional motion segmentation problems we find the majority of them addressing trajectories of feature points from independent 3-D rigid objects [5, 7, 13, 16, 26, 33] and the problem eventually boils down to subspace clustering. Other works also exploit dependant/articulated rigid body motion [36] or a motion model leading to nonlinear manifold clustering [32]. The group motion segmentation problem considered here has little in common with them. On the other hand, the non-rigid Structure-from-Motion problems [4, 3, 34] assume non-rigid shape to be linear combination of rigid ones, and non-rigid motion segmentation [35] makes use of local piecewise subspace model, while the group motion under our consideration does not belong to either of these cases.

In this work we employ Lie group theory [27] and in particular establish a statistical model over Lie algebra. Lie group and Lie algebra based approaches play roles in invariant visual modeling and recognition [9, 25], robotics [6], 3-D rigid motion estimation [10, 1, 30], as well as dense flow field modeling [20]. In this work, we discuss a new application to group motion estimation.

The proposed model is detailed in Section 2, and its application to group motion segmentation is presented in Sections 3 and 4. Section 5 empirically demonstrates the application of the approach, followed by a discussion in Section 6.

2. Spatio-Temporal Driving Force Model for A Group Motion Pattern

In this section we introduce a characterization for a group motion pattern, made of a collection of spatio-temporal constrained trajectories possibly noisy, of varying number, and undergoing spatio-temporal variation from realization to realization. The key idea is that we model the group motion as a dynamic process driven by a *spatio-temporal driving force* densely distributed across the area where the motion occurs, instead of simply as a set of discrete point trajectories. To be precise, the driving force is denoted as a 3×3 real matrix $F(t_0, t_f, x, y)$ which moves an object located at $X(t_0) = (x(t_0), y(t_0), 1)^T$ in homogeneous coordinates at time t_0 to location $X(t_f) = (x(t_f), y(t_f), 1)^T$ at time t_f, by

$$X(t_f) = F(t_0, t_f, x, y)X(t_0). \tag{1}$$

Without loss of generality, we usually take $t_0 = 0$ to be the starting time. It is obvious that once we have learned F for all t_f, x, and y, then the group motion is completely characterized. To be able to learn F, we limit our attention to those parametric F's which have the following properties: 1) $F(t_1, t_2, x, y)F(t_2, t_3, x, y) = F(t_1, t_3, x, y)$; 2) $F(t_1, t_2, x, y)^{-1} = F(t_2, t_1, x, y)$; and 3)

$$F(t, t+1, x, y) \triangleq F(t, x, y) = \begin{bmatrix} F_{11}(t,x,y) & F_{12}(t,x,y) & F_{13}(t,x,y) \\ F_{21}(t,x,y) & F_{22}(t,x,y) & F_{23}(t,x,y) \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

The $F(t, x, y)$'s defined this way is in fact a Lie group and more specifically an affine group [27]. By making use of Lie group theory we may achieve both generality and flexibility in modeling complex motion patterns, as shown next.

$F(t, x, y)$ characterizes the motion potential at time t for an object located at (x, y). However, we may look into an alternative representation. Consider $F(t, t + \delta t, x, y)$ and $X(t + \delta t) = F(t, t + \delta t, x, y)X(t)$, and we then have $X(t + \delta t) - X(t) = (F(t, t + \delta t, x, y) - I)X(t)$ where I is the identity matrix. Dividing both sides by δt and letting $\delta t \to 0$, we get $X'(t) = \mathbf{f}(t, x, y)X(t)$, in which

$X'(t) = (x'(t), y'(t), 0)^T$ is the speed of the object, and

$$\mathbf{f}(t,x,y) = \begin{bmatrix} f_{11}(t,x,y) & f_{12}(t,x,y) & f_{13}(t,x,y) \\ f_{21}(t,x,y) & f_{22}(t,x,y) & f_{23}(t,x,y) \\ 0 & 0 & 0 \end{bmatrix} \triangleq \lim_{\delta t \to 0} \begin{bmatrix} \frac{F_{11}(t,t+\delta t,x,y)-1}{\delta t} & \frac{F_{12}(t,t+\delta t,x,y)}{\delta t} & \frac{F_{13}(t,t+\delta t,x,y)}{\delta t} \\ \frac{F_{21}(t,t+\delta t,x,y)}{\delta t} & \frac{F_{22}(t,t+\delta t,x,y)-1}{\delta t} & \frac{F_{23}(t,t+\delta t,x,y)}{\delta t} \\ 0 & 0 & 0 \end{bmatrix}. \quad (3)$$

In fact, $\mathbf{f}(t,x,y)$ is the Lie algebraic representation of $F(t,x,y)$ and the two are related by the exponential map $F(t,x,y) = \exp(\mathbf{f}(t,x,y)) = \sum_{i=0}^{\infty} \frac{1}{i!}\mathbf{f}(t,x,y)^i$ and logarithmic map $\mathbf{f}(t,x,y) = \log(F(t,x,y)) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i}(F(t,x,y)-I)^i$ [27]. In other words, we may work equivalently with $\mathbf{f}$ instead of directly working with F for the driving force model. The advantage of using $\mathbf{f}$ lies in the fact that the space of all $\mathbf{f}$'s, the Lie algebra, is a linear one on which we may develop various statistical tools, while the Lie group of F's is a nonlinear manifold not easy to work with. More importantly, $X'(t) = \mathbf{f}(t,x,y)X(t)$ implies that the location and speed of the object are linearly related by $\mathbf{f}$, and both the location and the speed are low-level features obtainable from the motion trajectories of the objects, *i.e.*, learning a single driving force model $\mathbf{f}(t,x,y)$ will be straightforward.

2.1. Learning a Spatial Hybrid Driving Force Model at a Time Instant

Suppose we fix a specific time instant t. Intuitively, different driving forces $\mathbf{f}(t,x,y)$'s induce different 'affine' motions for different (x,y)'s, and learning for all (x,y)'s in the whole area of group motion is intractable. On the other hand, constant $\mathbf{f}(t,x,y) \equiv \mathbf{f}(t)$ induces a global 'affine' motion to all objects in the group, whose representative power is severely limited. For this reason, we propose a *spatial hybrid* model, in which we assume K driving forces in the area of group motion. The effective area of the kth ($k = 1, 2, \cdots, K$) force is Gaussian distributed as $(x,y)^T \sim \mathcal{N}(\mu^k, \mathbf{\Sigma}^k)$. (We drop the argument t in this subsection for simplicity.) In the effective area, there is a uniform 'affine' motion $\mathbf{f}^k$. For notational convenience we write

$$\mathbf{f}^k = \begin{bmatrix} A^k & \mathbf{b}^k \\ \mathbf{0}^T & 0 \end{bmatrix} \quad (4)$$

where A^k is the upper-left 2×2 block of $\mathbf{f}^k$.

Consider the feature vector $Y \triangleq (x, y, x', y')^T$ extracted from an object motion trajectory driven by the kth force, and it is obvious that

$$Y = \begin{bmatrix} I \\ A^k \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{b}^k \end{bmatrix}. \quad (5)$$

Taking into account the noise created due to approximating the speed (1st order derivative) of point trajectories, we represent the observed feature vector as

$$\mathbf{y} = Y + \begin{bmatrix} \mathbf{0} \\ \mathbf{n}^k \end{bmatrix} \quad (6)$$

where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, T^k)$. Then after some manipulations we find $\mathbf{y} \sim \mathcal{N}(\nu^k, \mathbf{\Gamma}^k)$, where

$$\nu^k = \begin{bmatrix} \mu^k \\ A^k\mu^k + \mathbf{b}^k \end{bmatrix} \quad (7)$$

and

$$\mathbf{\Gamma}^k = \begin{bmatrix} \mathbf{\Sigma}^k & \mathbf{\Sigma}^k A^{kT} \\ A^k\mathbf{\Sigma}^k & A^k\mathbf{\Sigma}^k A^{kT} + T^k \end{bmatrix}. \quad (8)$$

We are now in a position to learn the spatial hybrid driving force model for a particular time instant t, from a limited discrete number of objects in a group motion. Suppose we have observed t_M motion feature vectors $\{\mathbf{y}_m\}_{m=1}^{t_M}$, then the learning task boils down to fitting a Gaussian Mixture Model (GMM) of K component $\{(\nu^k, \mathbf{\Gamma}^k)\}_{k=1}^K$ (using uniform mixing coefficient $\frac{1}{K}$). An Expectation-Maximization procedure is employed to complete the inference. After successfully learning the GMM, the spatial hybrid driving model, *i.e.*, $\{(\mathbf{f}^k, \mu^k, \mathbf{\Sigma}^k)\}_{k=1}^K$, is recovered. Note that by Gaussian assumption the learned effective areas of different driving force components will overlap and cover the whole area; to eliminate this ambiguity we technically partition the area such that the distance between the component's center and (x,y) is minimized.

It may be useful to recap what has been achieved till now by using the spatial hybrid driving force model. In fact, at time t we partition the area into K subareas, and model the instant motions of all the objects in one subarea to be an uniform 'affine' motion. In this way we are actually establishing a dense 'motion potential field' across the area where the group motion happens. Though the field may be learned from motion features of sparse objects, it exists everywhere, and any other group of objects giving rise to the same field model is regarded as the same group motion pattern at time t, though participating objects (which the motion features come from) may appear in different locations from one group pattern to another. Figure 2 gives two examples of the spatial hybrid driving force model, where the sparse objects for learning the model, the learned partition of the area, as well as the learned driving force in each partition are all shown for two group motion samples in the dataset.

2.2. Learning Temporal Evolution of Driving Force Models for a Group Motion Pattern

Having obtained a hybrid driving force model at a particular time instant, we turn to the temporal evolution of the model, which eventually characterizes the complete group motion. Denote the driving force model at time t as $\mathfrak{M}(t) =$

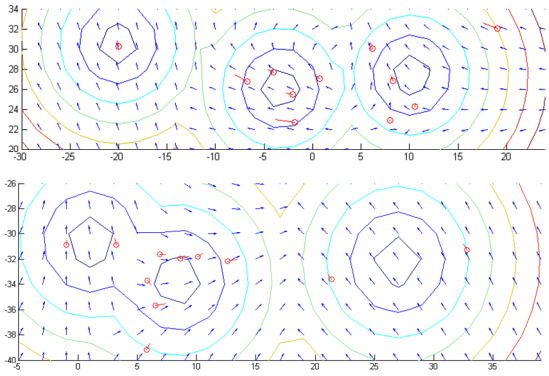


Figure 2. Two samples of 3-component driving force model at a time instant. The red circles denote the relative objects (offensive players) and the red bars attached to them denote the velocities of the objects. The blue arrow array gives the densely distributed driving force learned from the sparse object motion observations. The contour lines enclose the effective areas for the driving forces.

$\{\mathbf{m}^k(t)\}_{k=1}^K$ where $\mathbf{m}^k(t) = (\mathbf{f}^k(t), \mu^k(t), \mathbf{\Sigma}^k(t))$, and assume we have learned models at time $t = t_1, t_2, \cdots, t_T$ (which may not be continuous, but intermittent instants due to fragmented trajectories). We then learn a temporal sequence of these models in two steps: 1) component alignment between two consecutive instants, and 2) learning a parametric representation for the temporal sequence.

Component alignment is performed on $\mathfrak{M}(t_{i+1})$ with respect to aligned $\mathfrak{M}(t_i)$, starting from $\mathfrak{M}(t_2)$ with respect to $\mathfrak{M}(t_1)$. Mathematically, let the vector $(k_1, k_2, \cdots, k_K)^T$ denote the element-wise permutation of vector $(1, 2, \cdots, K)^T$, then we aim to find an optimal permutation such that $\sum_{j=1}^K D(\mathbf{m}^j(t_i), \mathbf{m}^{k_j}(t_{i+1}))$ is minimized, where $D(\mathbf{m}, \mathbf{m}')$ is a properly defined dissimilarity between model $\mathbf{m}$ and $\mathbf{m}'$. In other words, we are trying to associate each driving force component at time t_{i+1} uniquely with one at time t_i such that within each associated pair are they as similar as possible. We then give each component at t_{i+1} a new component index which is nothing but that of the associated component at t_i. Obviously, after component alignment for a fixed k, $\mathbf{m}^k(t)$ become similar, or change smoothly, among all t_i's, which reduces the complexity of the parametric representation.

Note that the driving force model $\mathbf{m}$ includes the 'force' term $\mathbf{f}$ and the effective area $\mathcal{N}(\mu, \mathbf{\Sigma})$, and $\mathbf{f}$ lies on the Lie algebra which is a linear space. Therefore, we use

$$D(\mathbf{m}, \mathbf{m}') = \|\mathbf{f} - \mathbf{f}'\| + \alpha(KL(\mathcal{N}(\mu, \mathbf{\Sigma})\|\mathcal{N}(\mu', \mathbf{\Sigma}')) + KL(\mathcal{N}(\mu', \mathbf{\Sigma}')\|\mathcal{N}(\mu, \mathbf{\Sigma}))) \quad (9)$$

where $KL(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. Then the optimal permutation can be solved using the classical Hungarian assignment algorithm [18].

Now we are looking for a parametric representation of the temporal sequence $\mathfrak{M}(t)$ for $t = 1, 2, \cdots$. Note that $\mathbf{m}^k(t)$ is essentially composed of $\mathbf{f}^k(t)$, the space of which is a Lie algebra (denoted as $\mathfrak{F}$), and $\mathcal{N}(\mu^k(t), \mathbf{\Sigma}^k(t))$, which is on the nonlinear manifold of 2×2 Gaussian's. As the nonlinearity brings analytical difficulty, we work with the parameters of $\mathcal{N}(\mu^k(t), \mathbf{\Sigma}^k(t))$, *i.e.*, $[\mu_1^k(t), \mu_2^k(t), \sigma_{11}^k(t), \sigma_{12}^k(t)(= \sigma_{21}^k(t)), \sigma_{22}^k(t)]^T \triangleq \mathbf{g}^k(t)$, rather than the Gaussian itself, and regard the space of $\mathbf{g}^k(t)$'s (denoted as $\mathfrak{G}$) to be linear as well. (Though linearity does not rigorously hold, it is an effective approximation.)

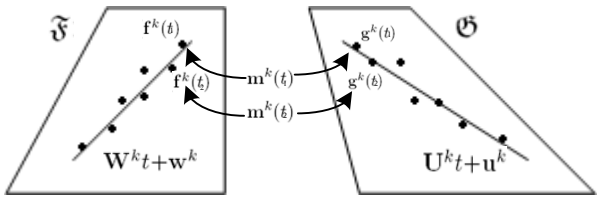


Figure 3. A pictorial illustration of the temporal evolution of driving force models. It only shows the kth component. In all, there should be K ones, *i.e.*, on left and right planes there should be K straight lines respectively.

We hence establish a parametric model for the temporal sequence $\mathfrak{M}(t), t = 1, 2, \cdots$ on the Cartesian product space $\mathfrak{F} \times \mathfrak{G}$. We propose the linear model $\{(\mathbf{W}^k t + \mathbf{w}^k + \mathbf{v}_1, \mathbf{U}^k t + \mathbf{u}^k + \mathbf{v}_2)\}_{k=1}^K$ for $\{(\mathbf{f}^k(t), \mathbf{g}^k(t))\}_{k=1}^K$, where

$$\mathbf{W}^k = \begin{bmatrix} W_{11}^k & W_{12}^k & W_{13}^k \\ W_{21}^k & W_{22}^k & W_{23}^k \\ 0 & 0 & 0 \end{bmatrix}, \quad (10)$$

$\mathbf{U}^k = [U_1^k, U_2^k, U_3^k, U_4^k, U_5^k]^T$, and $\mathbf{v}_1, \mathbf{v}_2$ are independent white Gaussian perturbation. With this model, $\mathbf{f}^k(t)$'s (resp. $\mathbf{g}^k(t)$'s) for each component k, when t varies, will approximately move in the one-dimensional subspace of $\mathfrak{F}$ (resp. $\mathfrak{G}$). In other words, components of the time-varying spatial hybrid driving force will evolve along straight lines in $\mathfrak{F} \times \mathfrak{G}$. A visual illustration for this idea is shown in Figure 3.

We use linear representations for the temporal sequence of driving forces basically for simplicity and effectiveness (to be demonstrated in the experiment). We may attempt advanced techniques but in this initial work on group motion segmentation we use linear ones to begin with. The straight lines $\{(\mathbf{W}^k t + \mathbf{w}^k, \mathbf{U}^k t + \mathbf{u}^k)\}_{k=1}^K$ in $\mathfrak{F} \times \mathfrak{G}$ are simply fitted using previously obtained models $(\mathbf{f}^k(t), \mathbf{g}^k(t))$'s at time $t = t_1, t_2, \cdots, t_T$ in the least square sense. However, once these lines are available, we may re-sample the lines at other t's to generate new spatial hybrid driving forces at those time instants. In this way, we actually learn the group motion pattern within the whole time duration from only motion information at limited time instants.

Up to now, a group motion pattern has been fully captured by $\{(\mathbf{W}^k t + \mathbf{w}^k, \mathbf{U}^k t + \mathbf{u}^k)\}_{k=1}^K \triangleq GM$, by which the motions of participants of the group activity, at any location and any time, are condensed into and may be recovered from the corresponding GM.

3. DP-DFM: Accounting for Group Motion Variation

The variation of group motion patterns from video to video leads to the variation of GM's learned from video to video. To statistically model this variability among different GM's, we establish a Dirichlet Process (DP) [2] over $\mathfrak{F} \times \mathfrak{G}$, leading to a Dirichlet Process - Driving Force Model (DP-DFM). The DP-DFM is essentially a Bayesian mixture model good for handling an unknown number of mixing components. As we do not have prior knowledge about the variability of the group motion patterns (*i.e.*, GM's from different offensive plays), DP-DFM is a natural choice.

Specifically, we regard the GM as a long vector consisting of the elements of $\mathbf{W}^k, \mathbf{w}^k, \mathbf{U}^k, \mathbf{u}^k, k = 1, \cdots, K$, and suppose it comes in the following manner (called 'Stick Breaking'): 1) Let $v_t \sim Beta(1, \eta)$ and $\lambda_t = v_t \Pi_{l=1}^{t-1}(1 - v_l)$; 2) Draw a sample from $\Sigma_{t=1}^{\infty} \lambda_t \delta(\theta_t)$, where $\delta(\theta_t)$ is a point measure situated at parameter vector θ_t, and $\theta_t \sim \mathcal{G}_0$, which is a base measure (Gaussian-Wishart in this work); 3) Draw a GM from a Gaussian whose mean and covariance are specified by θ_t. In this way the DP-DFM formulation has become a canonical DP mixture problem and we employ the standard procedure [2] to complete the inference of DP-DFM.

4. Probabilistic Segmentation

With a DP-DFM learned from training group motion patterns, we perform segmentation on a new testing motion pattern by synthesizing a set of Monte Carlo samples (*i.e.*, spatio-temporal DFM's) from DP-DFM, matching the trajectories in the testing motion pattern with these simulated models, and voting for the best matched motion trajectories as the segmentation result. To simulate a Monte Carlo sample, we first draw a GM from the DP-DFM. Then we recover the temporal sequence of spatially hybrid driving forces $\mathfrak{M}(t) = \{\mathbf{m}^k(t)\}_{k=1}^K$, and consequently the time-varying densely distributed driving forces $F(t, x, y)$. As a result, at each time instant t and for each object(trajectory) at t in the testing motion pattern, we may predict its location at $t + 1$ by (1), and measure the discrepancy between the predicted and actual locations (The discrepancy is simply measured as the distance between the two in this work). Those objects(trajectories) which accumulatively have the least discrepancies across all t's with all simulated driving force samples are finally determined as the participating objects.

5. Experiments

5.1. GaTech Football Play Dataset

We perform group motion segmentation on the GaTech Football Play Dataset, dividing players into participating (offensive) ones and non-participating (defensive) ones solely by their motion trajectories. The GaTech Football Play Dataset is a newly established dataset for group motion pattern/activity modeling and analysis. Recent works using this dataset [19, 29] reported results on play strategy recognition. The dataset consists of a collection of 155 NCAA football game videos (of which 56 are now available). Each video is a temporally segmented recording of an offensive play, *i.e.*, the play starts at the beginning of the video and terminates at the end. Accompanied with each video is a ground-truth annotation of the object locations in each frame. The annotation includes coordinates in the image plane of all the 22 players as well as field landmarks - the intersections of the field lines. Using the landmark information we can compute the plane homographies and convert motion coordinates into ground plane trajectories. We show the ground truth trajectories (in the ground plane coordinates) of sample plays in Figure 1.

We perform three rounds of experiments, the first of which employs ground-truth trajectories from training to testing. As in any practical system the input trajectories will be noisy, in the second round we generate noisy trajectories from the ground-truth and experiment with them. In the final round we test the learned framework on tracks computed from videos with a state-of-art multi-object tracker. In each round of experiments, we carry out multiple passes of five-fold evaluations, *i.e.*, in each pass we randomly divide $4/5$ of the samples into the training set and the remaining samples into testing set. The final statistics is aggregated from the average of all passes. Empirically we find $K = 5$ is a good selection for the total number of components. For Gaussian-Wishart prior $\mathcal{G}_0$, we set the Wishart scale matrix to be the inverse of sample covariance of training GM's, and Gaussian mean to be the sample mean of training GM's. The other free parameters in the framework are determined by experimental evaluation.

5.2. Experiment on Ground-Truth Trajectories

In the experiment with ground-truth trajectories, we have approximately $56 \times 4/5$ group motion samples, which may be captured in different views. For convenience and without loss of generality, we apply a homographic transform to each of them to work in a canonical view (ground plane in this work). To get sufficient exemplars to train the DP-DFM, we augment the training sets by generating new training samples from the original ones. For this purpose, we perturb each original trajectory by adding 2-D isotropic Gaussian on ground-plane coordinates at multiples of 20% of the whole motion duration, and polynomially interpolat-

Table 1. The segmentation rates comparison (%).

Proposed Driving Force Model	79.7
Homogeneous Spatial Model	74.8
Time-Invariant Model(similar to [20])	73.3
Linear Time-Invariant System Model([22])	70.7

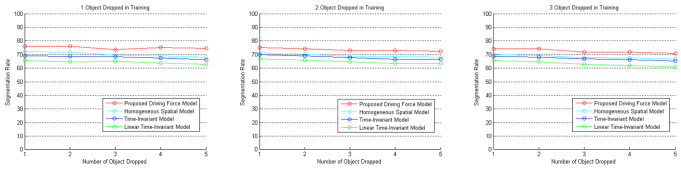


Figure 5. Segmentation statistics on non-robust trajectories.

ing the other time instants. In this way, we generate 25 new motion patterns from each original one. When learning a single hybrid driving force model within each (original or generated) motion pattern, we perform discriminative training, *i.e.*, we not only collect location/speed pairs from relevant (offensive) trajectories, but also take into account the irrelevant (defensive) trajectories away from the relevant ones, and include the inverse speeds from them into consideration. In addition, each speed vector is replicated a couple of times in the neighborhood of the corresponding location.

For comparison we set up three baselines. The first uses the homogeneous spatial model, *i.e.*, $K = 1$, and the second uses the time-invariant model, *i.e.*, we use a fixed hybrid driving force for all t's. Note that the second baseline is in principle similar to the model in [20]. In the third baseline, we simply regard the relevant trajectories as noisy observations of the states of a linear time-invariant dynamic system and use standard system identification approach [22] to learn the model.

We use the ratio of the correct segmented offensive players to the total offensive players, namely segmentation rate, as the numerical criterion to evaluate the performance, which is shown in Table 1. Samples of the segmentation results, are shown in Figure 4.

5.3. Experiment on Non-robust Trajectories

In this experiment, we simulate non-robustness by first randomly dropping a few trajectories (1, 2, 3 when training and 1, 2, 3, 4, 5 when testing) from the ground-truth, and then for the remaining trajectories randomly locating the durations during which we let the trajectories disappear (using a 1/0 birth-death process model with $\lambda/\mu = 5$). The training samples are then augmented by perturbing trajectories in every continuous durations. The statistics are shown in Figure 5. It turns out that the segmentation performance is insensitive to the varying number of missing trajectories as well as interruptions, as expected from having a dense field and continuous sequence.

5.4. Experiment on Trajectories from Tracking

In this evaluation, we employ a multi-object tracker [15] rather than directly using the annotations. As before, the trajectories are then transformed into ground plane coordinates. The multi-object tracking algorithm is based on foreground detection and tends to merge multiple targets into a single one (thus loses objects) when objects are small, highly cluttered, or strongly occluded. Note that in this case no numerical statistics can be calculated due to difficulty in associating these non-robust tracks with the ground-truth. However, we show the results in Figure 6 for a qualitative demonstration of the performance.

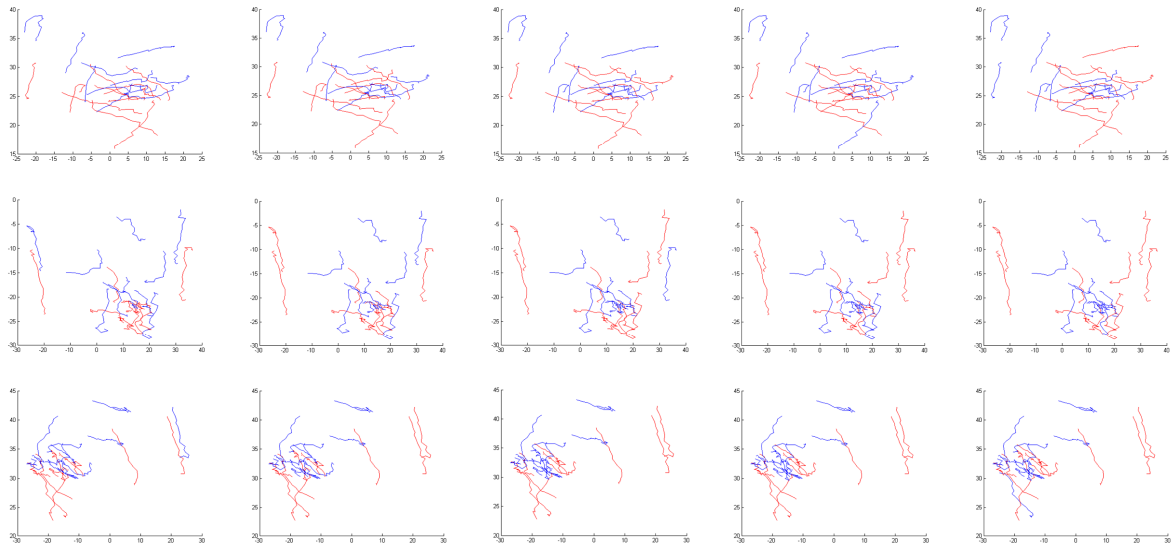
6. Discussions

We briefly discuss a few related issues. The first is the fact that the group motion segmentation algorithm can be used for temporal detection of the group motion, *i.e.*, to determine the starting and ending location along the time axis. As the GaTech Football Play Dataset only provides temporally segmented video clips containing a complete play, we are unable to empirically show this. However, for this purpose we simply initialize the segmentation algorithm from different time instants and identify the one(s) with the most likely match(es). Note that the algorithms can run in parallel.

A second issue is about estimating the spatial area of the group motion pattern. In football plays the participants and non-participants are homogeneously mixed all across the whole area of interest. However, in other applications the group motion pattern may only occupy a small portion of the whole area of view. In this case, we re-scale the field model into multiple scales and run the algorithm in parallel and in multiple scales. Within each scale we run the algorithm in dense partitions of the whole field. Note that scales and partitions with low matches in early stage can be eliminated from the candidate pool and the computational cost will keep decreasing.

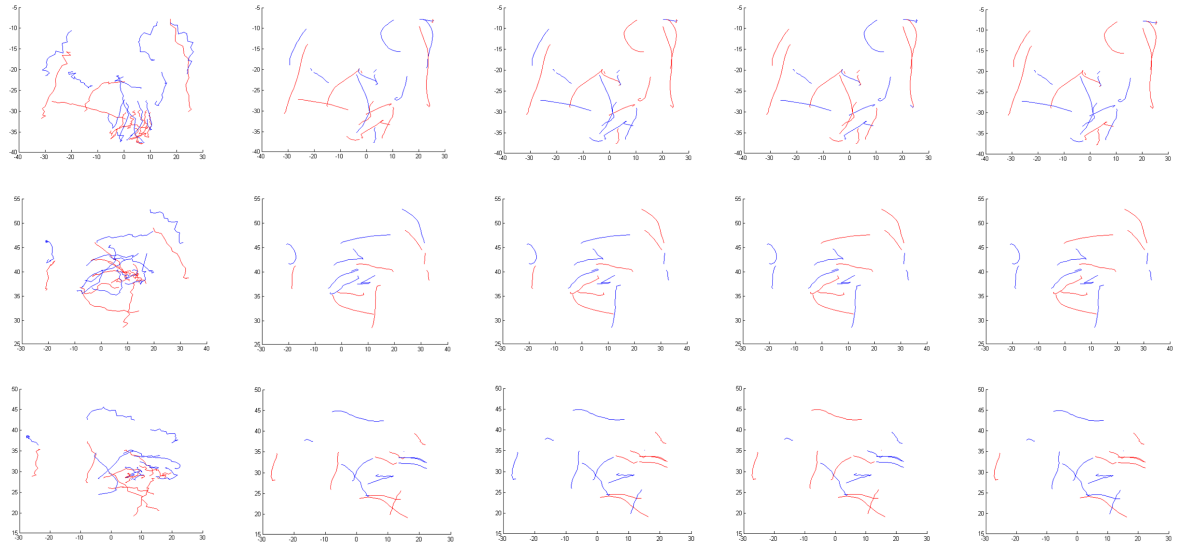
While the football play involves only one participating group, the method we presented can be extended to scenarios with multiple groups without much effort. To do so we learn a DFM per group and the testing motion pattern will be matched against every model. To get the final segmentation we simply vote for the best match.

We use the idea of learning a dense field model using sparse motions and testing sparse motions using dense field model. However, the model can be generalized to applications with dense motions. The dense trajectories, *e.g.*, may come from temporal association of SIFT or other local features of interesting points across frames, leading to local feature trajectories. Consequently, the model can be expected to work on problems regarding crowd in public areas, articulated human actions, as well as layered, moving and dynamic textures.



(a) Ground truth (b) Proposed DFM (c) Homogeneous Spatial model (d) Time-invariant model (similar to [21]) (e) Linear time-invariant system ([23])

Figure 4. Samples of segmentation results. In each row are a ground-truth group motion and corresponding segmentation results. Red trajectories denote the relevant objects and blue ones are irrelevant ones.



(a) Ground truth (b) Proposed DFM (c) Homogeneous Spatial model (d) Time-invariant model (similar to [21]) (e) Linear time-invariant system ([23])

Figure 6. Samples of segmentation results on trajectories from tracking. In each row are a ground-truth group motion and corresponding segmentation results on tracks. Red trajectories denote the relevant objects and blue ones are irrelevant ones.

The model is not view-invariant. We need to learn a separate model for each static view. However, static cameras are typical for surveillance and also commonly used for sports recordings. Also, the synthesis and voting based method is not computationally economic, and thus needs further improvement.

A final point is that though we designed methods in the context of group motion segmentation, the learned model, or compact features derived from it, can potentially serve as representatives of the underlying group motion. This implies a possibility that the proposed framework can be used toward the original motivating application - group activity recognition. Meanwhile, it is also expected that integration of the model into a multi-object tracker will help to improve

the tracking quality due to its capability to predict potential motion. These open issues are under our further investigation.

Acknowledgement: The authors thank Sima Taheri, Dr. Shaohua Kevin Zhou, and Minhua Chen for assistances and discussions. The authors also thank GaTech Athletic Association and GaTech School of Interactive Computing for providing the dataset. This work was funded by the VIRAT program.

References

[1] E. Bayro-Corrochano and J. Ortegon-Aguilar. Lie algebra approach for tracking and 3d motion estimation using monocular vision. *Image and Vision Computing*, 25:907 – 921, 2007. 2

[2] D. Blei and M. Jordan. Variational inference for dirichlet process mixtures. *Journal of Bayesian Analysis*, 1:121 – 144, 2006. 5

[3] M. Brand. Morphable 3d models from video. In *CVPR*, 2001. 2

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000. 2

[5] J. Costeira and T. Kande. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29:159 – 179, 1998. 1, 2

[6] T. Drummond and R. Cipolla. Application of lie algebras to visual servoing. *International Journal of Computer Vision*, 37:21 – 41, 2000. 2

[7] C. Gear. Multibody grouping from motion images. *International Journal of Computer Vision*, (29):133 – 150, 1998. 1, 2

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

[9] L. V. Gool, T. Moons, E. Pauwels, and A. Oosterlinck. Vision and lies approach to invariance. *Image and Vision Computing*, 13:259 – 277, 1995. 2

[10] V. M. Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *CVPR*, 2004. 2

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

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

[13] N. Ichimura. Motion segmentation based on factorization method and discriminant criterion. In *ICCV*, 1999. 1, 2

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

[15] S. Joo and R. Chellappa. A multiple-hypothesis approach for multiobject visual tracking. *IEEE Transactions on Image Processing*, 16(11):2849 – 2854, 2007. 6

[16] K. Kanatani. Motion segmentation by subspace separation and model selection. In *ICCV*, 2001. 1, 2

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

[18] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, (2):83 – 97, 1955. 4

[19] R. Li, R. Chellappa, and S. Zhou. Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition. In *CVPR*, 2009. 1, 5

[20] D. Lin, E. Grimson, and J. Fisher. Learning visual flows: A lie algebraic approach. In *CVPR*, 2009. 2, 6

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

[22] L. Ljung. *System Identification - Theory for the User*. Prentice Hall, 1999. 6

[23] X. Ma, F. Bashir, A. Khokhar, and D. Schonfeld. Event analysis based on multiple interactive motion trajectories. *IEEE Transactions on Circuits and Systems for Video Technology*, 19:397 – 406, 2009. 1

[24] B. Ni, S. Yan, and A. Kassim. Recognizing human group activities by localized causalities. In *CVPR*, 2009. 1

[25] R. P. N. Rao and D. L. Ruderman. Learning lie groups for invariant visual perception. In *NIPS*, 1999. 2

[26] S. R. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *CVPR*, 2008. 1, 2

[27] W. Rossmann. *Lie Groups: An Introduction through Linear Groups*. Oxford University Press, 2003. 2, 3

[28] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009. 1

[29] E. Swears and A. Hoogs. Learning and recognizing American football plays. In *Snowbird Learning Workshop*, 2009. 5

[30] O. Tuzel, R. Subbarao, and P. Meer. Simultaneous multiple 3d motion estimation via mode finding on lie groups. In *ICCV*, 2005. 2

[31] 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, October 2005. 1

[32] R. Vidal and Y. Ma. A unified algebraic approach to 2-d and 3-d motion segmentation. *Journal of Mathematical Imaging and Vision*, 25:403 – 421, 2006. 1, 2

[33] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1945 – 1959, 2005. 1, 2

[34] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. In *ECCV*, 2004. 2

[35] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *ECCV*, 2006. 1, 2

[36] L. Zelnik-Manor and M. Irani. Degeneracies, dependencies and their implications in multi-body and multi-sequence factorizations. In *CVPR*, 2003. 1, 2

[37] Y. Zhou, S. Yan, and T. S. Huang. Pair-activity classification by bi-trajectories analysis. In *CVPR*, 2008. 1