Parsing Collective Behaviors by Hierarchical Model with Varying Structure

Cong Zhang, Xiaokang Yang, Jun Zhu, Weiyao Lin
Institute of Image Communication and Network Engineering,
Shanghai Key Labs of Digital Media Processing and Communication
Shanghai Jiao Tong University, Shanghai 200240, China
{zhangcong0929, xkyang, junnyzhu, wylin}@sjtu.edu.cn

ABSTRACT

Collective behaviors are usually composed of several groups. Considering the interactions among groups, this paper presents a novel framework to parse collective behaviors for video surveillance applications. We first propose a latent hierarchical model (LHM) with varying structure to represent the behavior with multiple groups. Furthermore, we also propose a multi-layer-based (MLB) inference method, where a sample-based heuristic search (SHS) is introduced to infer the group affiliation. And latent SVM is adopted to learn our model. With the proposed LHM, not only are the collective behaviors detected effectively, but also the group affiliation in the collective behaviors is figured out. Experiment results demonstrate the effectiveness of the proposed framework.

Categories and Subject Descriptors

I.5 [**PATTERN RECOGNITION**] : [Models - *Structural*] and [Applications - *Computer vision*]

Keywords

Collective behavior, latent hierarchical model, sampling-based heuristic search, multi-layer-based inference method

1. INTRODUCTION

Recently, detecting human collective and group behaviors has attracted intensive research interests [5-7,11-13]. To detect collective behavior from surveillance videos is a challenge task. Some example collective behaviors include walking together, approaching, splitting, people being followed, etc. Generally, a collective behavior is composed by one or more groups [6], and the information of the interaction between groups will be helpful to detect the behaviors.

Many previous researches on group event detection consider the contextual information among the individuals for recognition. Zhou et al. [12] design a set of features such as causality ratio and GCT feedback ratio for describing the pair-activities. In [7], Ni et al. propose some types of localized causalities to characterize the local interaction of motion trajectories. However, most of these work mainly focus on the activity of only one group. Since collective behaviors are usually composed of multiple groups [6], the group-level activities will be inherently missed by these algorithms, as illustrated in Figure 1. Furthermore, although some generative-model-based algorithms [11] can deal with the multi-group problem by introducing some layered structure, most of these methods have the disadvantage of requiring large-scale training samples [8] and the low flexibility of handling varying number of people in the groups [6]. The most related work to ours is [5], where a discriminative latent model is used for recognizing group activities. However, this work still focuses on single-group recognition.

MM'12, October 29-November 2, 2012, Nara, Japan.

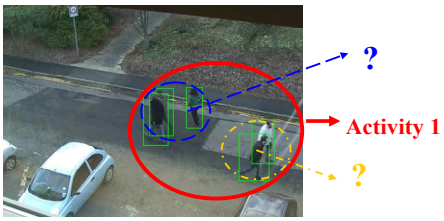


Figure 1. Example of a collective group behavior: without considering the group-level contextual information, it is difficult to detect the multi-group behaviors.

In this paper, we propose a novel framework on collective behavior recognition. In this framework, a new latent hierarchical model (LHM) with varying structure is introduced to parse collective behaviors into many groups such that the interactions of both individuals and groups can be considered. Meanwhile, we also present a new multi-layer-based (MLB) method to infer the LHM, and it can effectively solve the grouping problem resulted from the addition of the group layer. Our framework has the following three advantages:

(a) The LHM associates activities of different levels into unified graphical model, and enables us to suitably handle the collective behavior with multiple groups.
(b) With the proposed MLB inference method, our model can also automatically figure out the group affiliations in the activities. (i.e., which group does each person belong to, and what is the group activity for each group)
(c) Since we allow the number of nodes in our LHM to change, the LHM also has the flexibility to recognize behaviors with varying number of people and varying number of groups.

The remainder of this paper is organized as follows: the structure and the details of the LHM are presented in section 2. In section 3, we discuss our MLB inference method and learning algorithm for the LHM. Section 4 shows the experimental results and section 5 concludes the paper.

2. THE LATENT HIERARCHICAL MODEL

In this section, we describe the proposed latent hierarchical model for recognizing collaborative behaviors. Our goal is to learn a hierarchical model that jointly captures the collective behavior among groups, the activity for each group, the individual person's action, and the interactions among them.

The graph structure of our LHM is shown in Figure 2. There are four levels contained in our model: the collective behavior level, the group activity level, the individual action level, and the image level. The collective behavior level is composed of one or more groups in the group activity level. And in each group there are

several individual actions in the person action level. Finally, features are extracted for each person in the image level for performing recognition in the higher levels. Note that in our model, the image level is not only linked to the individual action level, rather, it also connects to the nodes in the other two levels such that more information can be utilized in the higher levels. It should also be noted that one of major contributions for our LHM is the introduction of the group activity level.

Inspired by [5], in order to represent the interaction among the nodes in the LHM, we propose six types of potentials: the Individual-Image Potential, the Individual-Individual Potential, the Group-Individual Potential, the Group-Group Potential, the Behavior-Group Potential, and the Group-Image Potential. They are described in Figure 2 and Table 1. In the following, we will describe how to use our LHM for detecting collective behaviors.

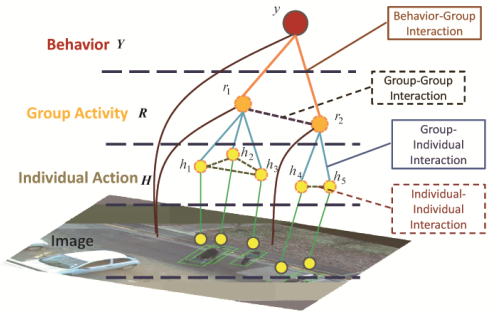


Figure 2. The structure illustration of the collective behavior model. The dashed lines and nodes represent that the interaction and the node labels are latent.

Let S be a video clip to be detected. We assume S contains M people and K groups. The extracted features for these people are denoted as $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2\cdots,\boldsymbol{x}_M\}$, where $\boldsymbol{x}_i$ is the feature for the i-th person. In our model, the individual action labels of all the people are denoted as $\boldsymbol{h}=(h_1,h_2,\cdots,h_M)$, where h_i is the action label of the i-th person. Similarly, the labels in the group activity level can be defined as $\boldsymbol{r}=(r_1,r_2,\cdots,r_K)$, where r_k is the group activity label of the k-th group. And the label in the collective behavior level is denoted as y. Let $\boldsymbol{H}$, $\boldsymbol{R}$, $\boldsymbol{Y}$ be the spaces of all possible label sets for the actions, group activities and collective behaviors respectively (i.e., $h\in\boldsymbol{H}$, $r\in\boldsymbol{R}$ and $y\in\boldsymbol{Y}$). The corresponding labels to the nodes in our LHM graphical model can be shown in Figure 2.

Based on our LHM, we construct a score function to evaluate the compatibility of the candidate labels and the graph structures. It is described by Eq. (1):

$$f_{\boldsymbol{\omega}}(y,\boldsymbol{h},\boldsymbol{r},\boldsymbol{X};\mathrm{G})=\boldsymbol{\omega}^T\cdot\Psi(y,\boldsymbol{h},\boldsymbol{r},\boldsymbol{X};\mathrm{G}) \tag{1}$$

where $f_{\boldsymbol{\omega}}(y,\boldsymbol{h},\boldsymbol{r},\boldsymbol{X};\mathrm{G})$ is the score value when the extracted image features are $\boldsymbol{X}$, the graph structure of the LHM is G, and the candidate labels for the individual action, group activity, and collective behavior are y, $\boldsymbol{h}$, and $\boldsymbol{r}$, respectively. Note that in our algorithm, the graph structure $\mathrm{G}=(\mathrm{V},\mathrm{E})$ contains nodes V and edges E. The graph is not fixed in order to handle the varying number of people and groups. That is, the number of person-action-level nodes M is dependent on the number of people in the current video clip while the number of group-activity-level nodes K is automatically inferred in our algorithm. The right part of Eq. (1) can be calculated by Eq. (2),

$$\begin{aligned}\boldsymbol{\omega}^T\cdot\Psi(y,\boldsymbol{h},\boldsymbol{r},\boldsymbol{X};\mathrm{G})=&\sum_{h_i\in\boldsymbol{H}}\boldsymbol{\omega}_1^T\phi_1(h_i,\boldsymbol{x}_i)+\sum_{k=1}^{K}\sum_{(h_i^k,h_j^k)\in\mathrm{E}}\boldsymbol{\omega}_2^T\phi_2(h_i^k,h_j^k,r_k)\\&+\sum_{k=1}^{K}\sum_{(r_k,h_i^k)\in\mathrm{E}}\boldsymbol{\omega}_3^T\phi_3(r_k,h_i^k)+\sum_{g_k\in\boldsymbol{R}}\boldsymbol{\omega}_5^T\phi_5(y,r_k)\\&+\sum_{(r_i,r_j)\in\mathrm{E}}\boldsymbol{\omega}_4^T\phi_4(r_i,r_j,y)+\sum_{r_k\in\boldsymbol{R}}\boldsymbol{\omega}_6^T\phi_6(r_k,\boldsymbol{x}_i)\end{aligned} \tag{2}$$

where $\boldsymbol{\omega}=\{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\boldsymbol{\omega}_3,\boldsymbol{\omega}_4,\boldsymbol{\omega}_5,\boldsymbol{\omega}_6\}$ is the concatenation of linear parameters for different potentials. And it is trained from the training set. $\Psi=(\phi_1,\phi_2,\phi_3,\phi_4,\phi_5,\phi_6)$ is the concatenated vector of all features given the features $\boldsymbol{X}$, candidate labels (y, $\boldsymbol{h}$, and $\boldsymbol{r}$), and the graph structure G. The calculation of these potentials is shown in Table 1.

Table 1. The definition of the six potentials (Here d_{kj} is the Euclidean distance between the person i and the center of the group k and 1(•) is the indicator function).

Feature Definition
Individual-Image Potential $\boldsymbol{\omega}_1^T\phi_1(h_i,\boldsymbol{x}_i)=\sum_{a\in H}\boldsymbol{\omega}_{1a}^T\mathbf{1}(h_i=a)\cdot\boldsymbol{x}_i$
Individual-Individual Potential $\boldsymbol{\omega}_2^T\phi_2(r_k,h_i,h_j)=\sum_{a\in\boldsymbol{R}}\sum_{b\in\boldsymbol{H}}\sum_{c\in\boldsymbol{H}}\boldsymbol{\omega}_{2abc}^T\cdot\mathbf{1}(\mathrm{r}_k=a)\cdot\mathbf{1}(h_i=b)\cdot\mathbf{1}(h_j=c)$
Group-Individual Potential $\boldsymbol{\omega}_3^T\phi_3(r_i,h_j)=\sum_{a\in\boldsymbol{R}}\sum_{b\in\boldsymbol{H}}\boldsymbol{\omega}_{3ab}^T\cdot\mathbf{1}(r_k=a)\cdot\mathbf{1}(h_j=b)\cdot(1,d_{kj})$
Group-Group Potential $\boldsymbol{\omega}_4^T\phi_4(r_i,r_j,y)=\sum_{a\in\boldsymbol{Y}}\sum_{b\in\boldsymbol{R}}\sum_{c\in\boldsymbol{R}}(\boldsymbol{\omega}_{4abc}^T\cdot\mathbf{1}(y=a)\cdot\mathbf{1}(r_i=b)\cdot\mathbf{1}(r_j=c))$
Behavior-Group Potential $\boldsymbol{\omega}_5^T\phi_5(y,r_k)=\sum_{a\in\boldsymbol{Y}}\sum_{b\in\boldsymbol{R}}\boldsymbol{\omega}_{5ab}^T\cdot\mathbf{1}(y=a)\cdot\mathbf{1}(r_k=b)\cdot(1,d_{yk})$
Group-Image Potential $\boldsymbol{\omega}_6^T\phi_6(r_k,\boldsymbol{x}_i)=\sum_{a\in\boldsymbol{R}}\boldsymbol{\omega}_{6a}^T\cdot\mathbf{1}(r_k=a)\cdot\boldsymbol{x}_i$

Therefore, based on the score function in Eq. (1), the recognition process with our LHM can be described by: given the input image feature $\boldsymbol{X}$ and a model parameters $\boldsymbol{\omega}$, try to find an optimal graph structure G and three optimal label sets such that $f_{\boldsymbol{\omega}}(y,\boldsymbol{h},\boldsymbol{r},\boldsymbol{X};\mathrm{G})$ is maximized:

$$(\mathrm{y}^*,\boldsymbol{h}^*,\boldsymbol{r}^*,\mathrm{G}^*)=\arg\max_{(\mathrm{y},\boldsymbol{h},\boldsymbol{r},\mathrm{G})}f_{\omega}(\mathrm{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{X};\mathrm{G}) \tag{3}$$

Then, the problem comes to how to learn a suitable parameter set $\boldsymbol{\omega}$ and how to infer the optimal labels and graph structure. And this will be solved by our MLB inference method. It will be described in the Sec 3.2.

3. LEARNING AND INFERENCE

3.1 Learning by latent SVM

Given a set of labeled training data $\langle\boldsymbol{X}^n,\boldsymbol{h}^n,y^n\rangle(n=1,2,\cdots,N)$, our goal is to learn the optimal model parameter $\boldsymbol{\omega}$. In order to achieve this, our goal is to let the ground truth labels y^n get a score higher than others. By this way, for a new input data $\boldsymbol{X'}$, we can not only classify $\boldsymbol{X'}$ to some behavior class y', but should also be able to find latent variables, the group activities $\boldsymbol{r'}$ as well as the group affiliation from the resulting graph structure G′. The latent SVM formulation [2] is adopted for learning as in Eq. (4).

$$\begin{aligned}&\min_{\boldsymbol{\omega}}\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{n=1}^{N}\xi_i\\&s.t.\ \max_{\mathrm{G}}\max_{\boldsymbol{r}}f_{\boldsymbol{\omega}}(\boldsymbol{X}^n,\boldsymbol{h}^n,\boldsymbol{r},y^n;\mathrm{G})-\max_{\mathrm{G}}\max_{\boldsymbol{r}}\max_{\boldsymbol{h}}f_{\boldsymbol{\omega}}(\boldsymbol{X}^n,\boldsymbol{h},\boldsymbol{r},y;\mathrm{G})\\&\geq\Delta(y,y^n)-\xi_i\ \ \forall n,\forall y\in\boldsymbol{Y}\\&\Delta(y,y^n)=\begin{cases}\partial y & \text{if } y\neq y^{\mathrm{n}}\\ 0 & otherwise\end{cases}\end{aligned} \tag{4}$$

where ∂y is the loss for predicting a different label from the ground truth y^n. In general, ∂y is set as 1. The process of learning will meet two challenges: (a) since the set of negative instances cannot be enumerable, solving Eq. (4) will be computationally intensive. (b) The optimization problem in Eq. (4) is non-convex because of the existence of latent variables. In order to solve the first challenge, cutting plane algorithm [10] is employed to take the place of the constraints in Eq. (4). In order

to solve the second challenge, we propose to introduce the non-convex bundle optimization process [3] for finding the cutting planes, where the cutting plane $c(\boldsymbol{\omega}) = a_{\boldsymbol{\omega}} \cdot \boldsymbol{\omega} + b_{\boldsymbol{\omega}}$ can be calculated by:

$$\begin{aligned} a_{\boldsymbol{\omega}} &= \sum_{n=1}^{N} (\Psi(\boldsymbol{X}^n, \boldsymbol{h}^*, \boldsymbol{r}^*, y^*; \mathrm{G}^*) - \Psi(\boldsymbol{X}^n, \boldsymbol{h}^n, \boldsymbol{r}^*, y^n; \mathrm{G}^*)) \\ b_{\boldsymbol{\omega}} &= \sum_{n=1}^{N} \max_{\mathrm{G}} \max_{y} \max_{\boldsymbol{r}} \max_{\boldsymbol{h}} f_{\boldsymbol{\omega}}(\boldsymbol{X}^n, \boldsymbol{h}, \boldsymbol{r}, y; \mathrm{G}) \\ &\quad + \Delta(y, y^n) - \max_{\mathrm{G}} \max_{\boldsymbol{r}} f_{\boldsymbol{\omega}}(\boldsymbol{X}^n, \boldsymbol{h}^n, \boldsymbol{r}, y^n; \mathrm{G}) \end{aligned} \tag{5}$$

According to Eq. (1) and (5), it is necessary to infer the optimal labels as well as the graph structure G. And the inference method will be described in the next subsection.

3.2 Inference by MLB

Given the model parameters $\boldsymbol{\omega}$, a set of labels (y, $\boldsymbol{h}$, $\boldsymbol{r}$) is scored by the function:

$$F_{\boldsymbol{\omega}}(\mathrm{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{X}; \mathrm{G}) = \max_{\boldsymbol{h}} \max_{\boldsymbol{r}} \max_{\mathrm{G}} \boldsymbol{\omega}^{\mathrm{T}} \cdot \Psi(y, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{X}; \mathrm{G}) \tag{6}$$

In order to find the highest score collective behavior label $y^* = \arg\max_y F_{\boldsymbol{\omega}}(y, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{X}; \mathrm{G})$, we should infer the optimal labels of the person action $\boldsymbol{h}$ and the group activities $\boldsymbol{r}$. Besides, the optimal graph structure G also needs to be inferred which contains the information of group affiliation as well as the individual action interactions in a group. As a combinatorial search, the optimization problem is NP-hard. Therefore, inspired by [5], we utilize an iterative two-step method to approximately solve this problem.

Step 1. Fix the graph structure G and optimize the action labels $\boldsymbol{h}$ and $\boldsymbol{r}$ for the set (y, $\boldsymbol{h}$, $\boldsymbol{r}$). In this step, the graph model G is divided into two homogeneous layers: the behavior-activity layer G_y and the activity-action layer G_r, as in Figure 3. Since these two layers can be inferred separately, step 1 can further be separated into two sub-steps:

(a) In the activity-action layer, first optimize the action labels $\boldsymbol{h}$ for all the possible group activity label sets in the label space $\boldsymbol{R}$.

$$\tilde{\boldsymbol{h}}_{\mathrm{q}}^* = \arg\max_{\boldsymbol{h}} \boldsymbol{\omega}^T \cdot \Psi(y, \boldsymbol{h}, \boldsymbol{r}_{\mathrm{q}}, \boldsymbol{X}; \mathrm{G}_r) \quad \forall \boldsymbol{r}_{\mathrm{q}} \in \boldsymbol{R} \tag{7}$$

(b) Then, in the behavior-activity layer, optimize the activity labels $\boldsymbol{r}$ for the all the possible collective behavior label values.

$$\tilde{\boldsymbol{r}}_p^* = \arg\max_{\boldsymbol{r}} \boldsymbol{\omega}^T \cdot \Psi(y_{\mathrm{p}}, \tilde{\boldsymbol{h}}_{\mathrm{r}}^*, \boldsymbol{r}, \boldsymbol{X}; \mathrm{G}_y) \quad \forall y_{\mathrm{p}} \in \boldsymbol{Y} \tag{8}$$

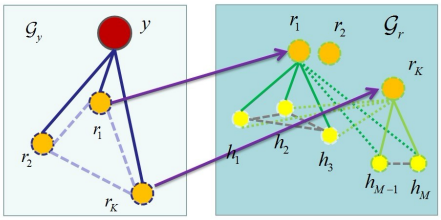


Figure 3. The two layers of graph G

Step 2. Fix all the labels and optimize graph structure G for the set (y, $\tilde{\boldsymbol{h}}_{\mathrm{q}}^*$, $\tilde{\boldsymbol{r}}_p^*$).

$$\tilde{\mathrm{G}}_{\mathrm{p}}^* = \arg\max_{\mathrm{G}} \boldsymbol{\omega}^T \cdot \Psi(y_{\mathrm{p}}, \tilde{\boldsymbol{h}}_{\mathrm{p}}^*, \tilde{\boldsymbol{r}}_p^*, \boldsymbol{X}; \mathrm{G}) \quad \forall y_{\mathrm{p}} \in \boldsymbol{Y} \tag{9}$$

In the above two steps, Step 1 is a standard max-inference problem in an undirected graph model. Thus, it can be solved by the loopy belief propagation (LBP) method [9]. For the second step, since the person action labels and the group activity labels are fixed, optimizing the graph structure in Eq. (9) is rewritten to:

$$\max_{\mathbf{Z}, \mathbf{L}} \sum_{k=1}^{K} \sum_{i,j} l_{ij}^k \psi_{i,j,k} + \sum_{k=1}^{K} \sum_{i=1}^{M} z_i^k \varphi_{i,k} \tag{10}$$

where $l_{ij}^k \in \mathbf{L}$ $\left(l_{ij}^k = \{0, 1\}\right)$ indicates whether the edge (i, j) is

Algorithm 1 The flow of inference

1: Input $(\boldsymbol{X}, y)$
2: Init: Find U search seeds $\mathrm{G}^{(1)}, \mathrm{G}^{(2)} \cdots, \mathrm{G}^{(U)}$
3: For every search seed, do
4: Hold G, infer the labels $\tilde{\boldsymbol{h}}^*, \tilde{\boldsymbol{r}}^*$
5: Hold $\tilde{\boldsymbol{h}}^*$ and $\tilde{\boldsymbol{r}}^*$, infer the graph structure $\tilde{\mathrm{G}}^*$
6: Until achieve convergence
7: Update the seeds, eliminate the lowest-score seed, $\mathrm{U} = \mathrm{U} - 1$
8: If $\mathrm{U} \neq 1$, return step 3
9: Else output the inference result $\boldsymbol{h}^*$, $\boldsymbol{r}^*$, G^*

indicates whether person i is in the k-th group (e.g., z_i^k=1 indicatesthat there is an edge linking person i's node in the person action level and group k's node in the group activity level). Let $\psi_{i,j,k}$ be the sum of all the pair-wise potentials in Table 1, if the pair of nodes (i, j) belonged to the group node k. And $\varphi_{i,k}$ be the sum of the unary potentials in Table 1, if the i-th person belongs to the k-th group.

The first term in Eq. (10) is to infer the interactions among people in the same group. And it can be solved by integer linear program (ILP) which adds an additional maximum-vertex-degree constraint to control the model complexity [5]. However, inferring the optimal group affiliations (i.e., the second term in Eq. (10)) is non-linear since the group-individual potential features are dependent on the group affiliation. If the number of people in the behavior is not large, all possible group affiliations are enumerable. However, since the grouping optimization problem is NP-hard, if the search space is too large, it tends to search out a local optimum from only one initial state.

In order to solve the optimization problem in Eq. (10), we propose a new sample-based heuristic search (SHS) approach to approximately infer the optimal group affiliation. The idea of the proposed SHS approach can be described in the following:

Firstly, U initial search seeds are drawn where each seed represents one candidate group affiliation. Then, for each group affiliation defined by search seed, the optimal label sets ($\tilde{\boldsymbol{h}}^*, \tilde{\boldsymbol{r}}^*$) as well as the optimal graph structure $\tilde{\mathrm{G}}^*$ are achieved by Eqs (7)-(10). Since there are U initial search seeds, we can achieve U label and graph structure sets. Then, each search seed is evaluated by calculating the compatibility score in Eq. (1) with the achieved ($\tilde{\boldsymbol{h}}^*, \tilde{\boldsymbol{r}}^*$ $\tilde{\mathrm{G}}^*$) sets and the lowest-score seed is eliminated. After that, each seed is updated to a different set of candidate group affiliations. Thus, we will have another U-1 new seeds. This process is iterated until converge or there is no seed left. Finally, the label and graph structure set that create the highest reliability score will be selected as the inference result.

Note that in the above process, the initial search seeds are calculated based on a prior distribution of the group number, which can be estimated from the training data (e.g., if the prior probability for having two groups is 0.7, then 70% of the initial seeds will have the group affiliation of two groups). Furthermore, since the object location information is intuitively the basic cue of group affiliation, the initial seeds are created based on the clustering results of individual location. More specifically, for each group number, we select the clustering result as well as its "closest" group states as the initial seeds under this group number. The "closeness" between seeds is evaluated by the group affiliation changing cost. That is, for each node v_i in the graph, $c_i = \min\{d_{\mathrm{G}_j}^i \ \forall j : v_i \notin \mathrm{G}_j\}$,where c_i is the cost of changing v_i and $d_{\mathrm{G}_j}^i$ is the Euclidean distance between v_i and the center of

group G_j not including v_i. Note that when updating the seeds in each iteration, the group affiliation changing cost is also utilized for selecting the nearest unused seed. To sum up, we summarize the flow of inference described algorithm as Algorithm 1.

4. EXPERIMENTS

In this section, we show experimental results for our proposed methods. Experiments are performed on the BEHAVE dataset [1] and try to detect four collective behaviors: *Approach* (AP), *Ingroup* (IG), *Split* (SP) and *Walk together* (WT). One example frame in this dataset is shown in Figure 1. Note that this experiment is challenging in that (a) some of the collective behaviors includes multiple groups, and (b) the number of people and groups in the collective behaviors are varying. Five long sequences are selected in our experiments with each sequence including 7000 to 10000 frames. And the experiments are performed by 50% training-50% testing protocol.

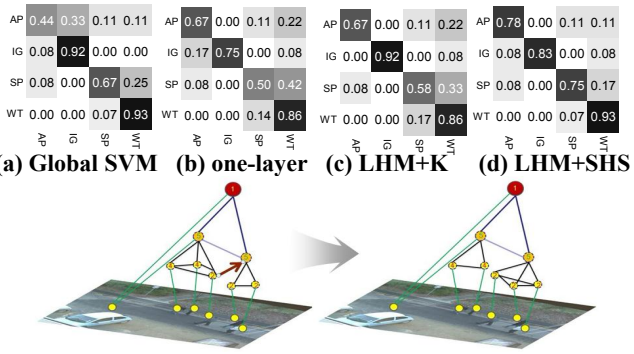
As mentioned, pedestrian detection is performed based on video clips which are obtained by a sliding window of length 50 and overlapping area 25. We extract five features for each person (i.e., $\boldsymbol{x_i}=[p_{ix}, p_{iy}, v_{ix}, v_{iy}, d_i]$ where (p_{ix}, p_{iy}) and (v_{ix}, v_{iy}) represent the location and speed of person i, and d_i represents the motion direction. In order to exclude the effect of tracking errors, all the features are derived from the ground-truth motion bound box (MBB). In practice, various tracking algorithms [4] can be used to extract the MBB information. Furthermore, we also set five activity types for the group activity level: standing, walk alone, run alone, walk together, and run together.

Four methods are compared in our experiments: (a) directly use linear SVM model classifier to recognize the input image-level features (global SVM in Table 2); (b) Use the one-layer model to represent the behavior [5] (one-layer in Table 2); (c) Use our LHM but directly use the K-means clustering (without SHS) to infer the group affiliations (LHM+K in Table 2); and (d) Use our LHM and SHS approach to infer the group affiliations (LHM+SHS in Table 2). Due to the limited space, only partial results are shown in this paper. Table 2 compares the total detection rate (TDR) (i.e., the behavior type will be determined the most types of video clip detection results) and the average per-class detection rate (ADR), which is the average value on the diagonal of the confusion matrix.

Table 2. Comparison of collective behavior classification accuracies for different methods.

	Global SVM	One-layer [5]	LHM+K	**LHM+SHS**
TDR	76.6	70.2	76.6	**83.0**
ADR	73.9	69.3	75.6	**82.3**

From Table 1 and Figure 4, it is clear that our LHM can achieve better results than the other methods. Specifically, our LHM has obviously stronger capability in handling the behaviors with multiple groups such as *Approach* and *Split*. Furthermore, it is noted that our LHM also has the advantage of being able to recognize the action for each individual person, the group activity for each group, and the group affiliations at one time. Compared to our model, the other methods are limited to only recognize the collective behavior or the individual actions. Finally, Figure 4 (e) shows the effectiveness of our SHS approach, if we directly use K-means clustering, the group affiliations may have low accuracy. However, with our SHS approach, we can have multiple seeds such that the correct group affiliations can be achieved. Besides, when the collective behavior is detected correctly, the precision of group affiliation is approximately 95%.



(a) Global SVM (b) one-layer (c) LHM+K (d) LHM+SHS

(e) The group affiliation of LHM+K(left) and LHM+SHS(right)

Figure 4. The confusion matrix(a-d) and the grouping states(e).

5. ACKNOWLEDGMENTS

This paper was supported in part by NSFC (61025005, 61129001, 61001146, 61071155), 973 Program (2010CB731401, 2010CB731406), and the 111 Project (B07022).

6. CONCLUSION

In this paper, we propose a novel framework to parse collective behaviors. In our framework, a latent hierarchical model with varying structure is introduced to handle the collective behavior with multiple groups. Furthermore, we also propose a multi-layer-based approach to precisely infer our LHM. Experimental results demonstrate the effectiveness of our method.

7. REFERENCES

[1] BEHAVE data: http://homepages.inf.ed.ac.uk/rbf/BEHAVE.

[2] Yang Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *Proc. CVPR*, pages 872 –879, June 2009.

[3] T. Do and T. Artières, "Large margin training for hidden markov models with partially observed states," *Int'l Conf. Machine Learning (ICML)*, pp. 265–272, 2009.

[4] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–574, 2003.

[5] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 99, pp. 1-14, 2011.

[6] W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Group event detection with a varying number of group members for video surveillance," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 20, no. 8, pp. 1057 – 067, 2010.

[7] B. Ni, S. Yan, and A. Kassim, "Recognizing human group activities with localized causalities," *IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 1470–1477, 2009.

[8] C. Wang, J. Zhang, J. Pu, X. Yuan and L. Wang, "Chrono-gait image: A novel temporal template for gait recognition," *Euro. Conf. Computer Vision (ECCV)*, pp. 257-270, 2010.

[9] J. Yedidia, W.T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," *Exploring artificial intelligence in the new millennium* (G. Lakemeyer and B. Nebel eds.), chapter 8, pp. 236–239, 2003.

[10] C. Yu and T. Joachims, "Learning structural SVMs with latent variables," *Proc. ICML*, pp. 1169–1176, 2009.

[11] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered HMMs," *IEEE Trans. Multimedia,* vol. 8, no. 3, pp. 509 –520, 2006.

[12] Y. Zhou, S. Yan, and T. Huang, "Pair-activity classification by bi-trajectories analysis," *Proc. CVPR* , pp. 1–8, 2008.