A Unified Framework for Multi-Target Tracking and Collective Activity Recognition

Wongun Choi and Silvio Savarese

Electrical and Computer Engineering, University of Michigan, Ann Arbor, USA
{wgchoi,silvio}@umich.edu

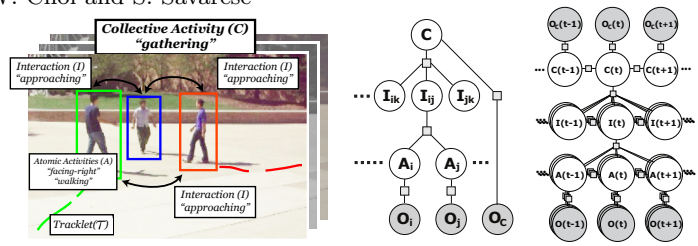
Abstract. We present a coherent, discriminative framework for simultaneously tracking multiple people and estimating their collective activities. Instead of treating the two problems separately, our model is grounded in the intuition that a strong correlation exists between a person's motion, their activity, and the motion and activities of other nearby people. Instead of directly linking the solutions to these two problems, we introduce a hierarchy of activity types that creates a natural progression that leads from a specific person's motion to the activity of the group as a whole. Our model is capable of jointly tracking multiple people, recognizing individual activities (*atomic activities*), the interactions between pairs of people (*interaction activities*), and finally the behavior of groups of people (*collective activities*). We also propose an algorithm for solving this otherwise intractable joint inference problem by combining belief propagation with a version of the branch and bound algorithm equipped with integer programming. Experimental results on challenging video datasets demonstrate our theoretical claims and indicate that our model achieves the best collective activity classification results to date.

Key words: Collective Activity Recognition, Tracking, Tracklet Association

1 Introduction

There are many degrees of granularity with which we can understand the behavior of people in video. We can detect and track the trajectory of a person, we can observe a person's pose and discover what *atomic activity* (*e.g.*, *walking*) they are performing, we can determine an *interaction activity* (*e.g.*, *approaching*) between two people, and we can identify the *collective activity* (*e.g.*, *gathering*) of a group of people. These different levels of activity are clearly not independent: if everybody in a scene is walking, and all possible pairs of people are approaching each other, it is very likely that they are engaged in a gathering activity. Likewise, a person who is gathering with other people is probably walking toward a central point of convergence, and this knowledge places useful constraints on our estimation of their spatio-temporal trajectory.

Regardless of the level of detail required for a particular application, a powerful activity recognition system will exploit the dependencies between different levels of activity. Such a system should reliably and accurately: (i) identify stable and coherent trajectories of individuals; (ii) estimate attributes, such as poses, and infer atomic activities; (iii) discover the interactions between individuals;



(a) (b) (c)

Fig. 1: In this work we aim at jointly and robustly tracking multiple targets and recognizing the activities that such targets are performing. **(a)**: The collective activity *"gathering"* is characterized as a collection of interactions (such as *"approaching"*) between individuals. Each interaction is described by pairs of atomic activities (e.g. *"facing-right"* and *"facing-left"*). Each atomic activity is associated with a spatial-temporal trajectory (tracklet τ). We advocate that high level activity understanding helps obtain more stable target trajectories. Likewise, robust trajectories enable more accurate activity understanding. **(b)**: The hierarchical relationship between atomic activities (A), interactions (I), and collective activity (C) in one time stamp is shown as a factor graph. Squares and circles represent the potential functions and variables, respectively. Observations are the tracklets associated with each individual along with their appearance properties O_i as well as crowd context descriptor O_c [1, 2] (Sec.3.1). **(c)**: A collective activity at each time stamp is represented as a collection of interactions within a temporal window. Interaction is correlated with a pair of atomic activities within specified temporal window (Sec.3.2). Non-shaded nodes are associated with variables that need to be estimated and shaded nodes are associated with observations.

(iv) recognize any collective activities present in the scene. Even if the goal is only to track individuals, this tracking can benefit from the scene's context. Even if the goal is only to characterize the behavior of a group of people, attention to pairwise interactions can help.

Much of the existing literature on activity recognition and tracking [3–11] avoids the complexity of this context-rich approach by seeking to solve the problems in isolation. We instead argue that tracking, track association, and the recognition of atomic activities, interactions, and group activities must be performed completely and coherently. In this paper we introduce a model that is both principled and solvable and that is the first to successfully bridge the gap between tracking and group activity recognition (Fig.1).

2 Related Work

Target tracking is one of the oldest problems in computer vision, but it is far from solved. Its difficulty is evidenced by the amount of active research that continues to the present. In difficult scenes, tracks are not complete, but are fragmented into tracklets. It is the task of the tracker to associate tracklets in order to assemble complete tracks. Tracks are often fragmented due to occlusions. Recent algorithms address this through the use of detection responses [12, 13], and pairwise interaction models [3–8]. The interaction models, however, are limited to a few hand-designed interactions, such as attraction and repulsion. Methods such as [14] leverage the consistency of the flow of crowds with models from physics, but do not attempt to associate tracklets or understand the actions of individuals. [15, 16] formulate the problem of multi-target tracking into a min-cost flow network based on linear/dynamic programming. Although both model interactions between people, they still rely on heuristics to guide the association process via higher level semantics.

A number of methods have recently been proposed for action recognition by extracting sparse features [17], correlated features [18], discovering hidden topic models [19], or feature mining [20]. These works consider only a single person,

and do not benefit from the contextual information available from recognizing interactions and activities. [21] models the pairwise interactions between people, but the model is limited to local motion features. Several works address the recognition of planned group activities in football videos by modelling the trajectories of people with Bayesian networks [9], temporal manifold structures [10], and non-stationary kernel hidden Markov models [22]. All these approaches, however, assume that the trajectories are available (known). In collective activity recognition, [23] recognizes group activities by considering local causality information from each track, each pair of tracks, and groups of tracks. [1] classifies collective activities by extracting descriptors from people and the surrounding area, and [2] extends it by learning the structure of the descriptor from data. [24] models a group activity as a stochastic collection of individual activities. None of these works exploit the contextual information provided by collective activities to help identify targets or classify atomic activities. [11] uses a hierarchical model to jointly classify the collective activities of all people in a scene, but they are restricted to modelling contextual information in a single frame, without seeking to solve the track identification problem. Finally, [25] recognizes the overall behavior of large crowds using a social force model, but does not seek to specify the behaviour of each individual.

Our contributions are four-fold: we propose (i) a model that merges for the first time the problems of collective activity recognition and multiple target tracking into a single coherent framework; (ii) a novel path selection algorithm that leverages target interactions for guiding the process of associating targets; (iii) a new hierarchical graphical model that encodes the correlation between activities at different levels of granularity; (iv) quantitative evaluation on a number of challenging datasets, showing superiority to the state-of-the-art.

3 Modelling Collective Activity

Our model accomplishes collective activity classification by simultaneously estimating the activity of a group of people (*collective activity* C), the pairwise relationships between individuals (*interactions activities* I), and the specific activities of each individual (*atomic activities* A) given a set of observations O (see Fig.1). A collective activity describes the overall behavior of a group of more than two people, such as *gathering*, *talking*, and *queuing*. Interaction activities model pairwise relationships between two people which can include *approaching*, *facing-each-other* and *walking-in-opposite-directions*. The atomic activity collects semantic attributes of a tracklet, such as poses (*facing-front*, *facing-left*) or actions (*walking*, *standing*). Feature observations $O = (O_1, O_2, ...O_N)$ operate at a low level, using tracklet-based features to inform the estimation of atomic activities. Collective activity estimation is helped by observations O_C, which use features such as spatio-temporal local descriptors [1, 2] to encode the flow of people around individuals. At this time, we assume that we are given a set of tracklets $\tau_1, ..., \tau_N$ that denote all targets' spatial location in 2D or 3D. These tracklets can be estimated using methods such as [6]. Tracklet associations are denoted by $T = (T_1, T_2, ..., T_M)$ and indicate the association of tracklets. We address the estimation of T in Sec.4.

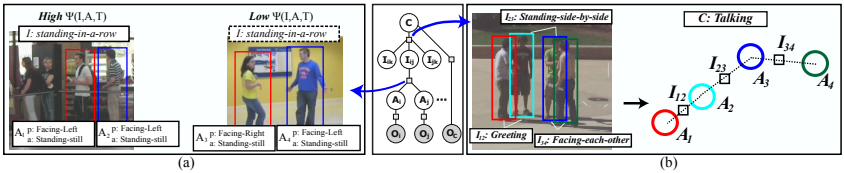


Fig. 2: **(a)**: Each interaction is represented by a number of atomic activities that are characterized by an action and pose label. For example, with interaction $I = \textit{standing-in-a-row}$, it is likely to observe two people with both $p = \textit{facing-left}$ and $a = \textit{standing-still}$, whereas it is less likely that one person has $p = \textit{facing-left}$ and the other $p = \textit{facing-right}$. **(b)**: Collective activity C is represented as a collection of interactions I. For example, with $C = \textit{talking}$ collective activity, it is likely to observe the interaction $I_{34} = \textit{facing-each-other}$, and $I_{23} = \textit{standing-side-by-side}$. The consistency of $C, I_{12}, I_{23}, I_{34}$ generates a high value for $\Psi(C, I)$.

The information extracted from tracklet-based observations O enables the recognition of atomic activities A, which assist the recognition of interaction activities I, which are used in the estimation of collective activities C. Concurrently, observations O_c provide evidence for recognizing C, which are used as contextual clues for identifying I, which provide context for estimating A. The bi-directional propagation of information makes it possible to classify C, A, and I robustly, which in turn provides strong constraints for improving tracklet association T. Given a video input, the hierarchical structure of our model is constructed dynamically. An atomic activity A_i is assigned to each tracklet τ_i (and observation O_i), an interaction variable I_{ij} is assigned to every pair of atomic activities that exist at the same time, and all interaction variables within a temporal window are associated with a collective activity C.

3.1 The model

The graphical model of our framework is shown in Fig.1. Let $O = (O_1, O_2, ...O_N)$ be the N observations (visual features within each tracklet) extracted from video V, where observation O_i captures appearance features $s_i(t)$, such as histograms of oriented gradients (HoG [26]), and spatio-temporal features $u_i(t)$, such as a bag of video words (BoV [17]). t corresponds to a specific time stamp within the set of frames $\mathcal{T}_V = (t_1, t_2, ..., t_Z)$ of video V, where Z is the total number of frames in V. Each observation O_i can be seen as a realization of the underlying atomic activity A_i of an individual. Let $A = (A_1, A_2, ..., A_N)$. A_i includes pose labels $p_i(t) \in \mathcal{P}$, and action class labels $a_i(t) \in \mathcal{A}$ at time $t \in \mathcal{T}_V$. $\mathcal{P}$ and $\mathcal{A}$ denote the set of all possible pose (e.g, *facing-front*) and action (e.g, *walking*) labels, respectively. $I = (I_{12}, I_{13}, ..., I_{N-1N})$ denotes the interactions between all possible (coexisting) pairs of A_i and A_j, where each $I_{ij} = (I_{ij}(t_1), ...I_{ij}(t_Z))$ and $I_{ij}(t) \in \mathcal{I}$ is the set of interaction labels such as *approaching*, *facing-each-other* and *standing-in-a-row*. Similarly, $C = (C(t_1), ..., C(t_Z))$ and $C(t_i) \in \mathcal{C}$ indicates the collective activity labels of the video V, where $\mathcal{C}$ is the set of collective activity labels, such as *gathering*, *queueing*, and *talking*. In this work, we assume there exists only one collective activity at a certain time frame. Extensions to modelling multiple collective activities will be addressed in the future. T describes the target (tracklet) associations in the scene as explained in Sec.3.

We formulate the classification problem in an energy maximization framework [27], with overall energy function $\Psi(C, I, A, O, T)$. The energy function is modelled as the linear product of model weights w and the feature vector ψ :

$$\Psi(C, I, A, O, T) = w^T \psi(C, I, A, O, T) \quad (1)$$

$\psi(C, I, A, O, T)$ is a vector composed of $\psi_1(\cdot), \psi_2(\cdot), ..., \psi_m(\cdot)$ where each feature element encodes local relationships between variables and w, which is learned discriminatively, is the set of model parameters. High energy potentials are associated with configurations of A and I that tend to co-occur in training videos with the same collective activity C. For instance, the *talking* collective activity tends to be characterized by interaction activities such as *greeting*, *facing-each-other* and *standing-side-by-side*, as shown in Fig.2.

3.2 Model characteristics

The central idea of our model is that the atomic activities of individuals are highly correlated with the overall collective activity, through the interactions between people. This hierarchy is illustrated in Fig.1. Assuming the conditional independence implied in our undirected graphical model, the overall energy function can be decomposed as a summation of seven local potentials: $\Psi(C, I)$, $\Psi(C, O)$, $\Psi(I, A, T)$, $\Psi(A, O)$, $\Psi(C)$, $\Psi(I)$, and $\Psi(A)$. The overall energy function can easily be represented as in Eq.1 by rearranging the potentials and concatenating the feature elements to construct the feature vector ψ. Each local potential corresponds to a node (in the case of unitary terms), an edge (in the case of pairwise terms), or a high order potential seen on the graph in Fig.1.(c): 1) $\Psi(C, I)$ encodes the correlation between collective activities and interactions (Fig.2.(b)). 2) $\Psi(I, A, T)$ models the correlation between interactions and atomic activities (Fig.2.(a)). 3) $\Psi(C)$, $\Psi(I)$ and $\Psi(A)$ encode the temporal smoothness prior in each of the variables. 4) $\Psi(C, O)$ and $\Psi(A, O)$ model the compatibility of the observations with the collective activity and atomic activities, respectively.

Collective - Interaction $\Psi(C, I)$: The function is formulated as a linear multi-class model [28]:

$$\Psi(C, I) = \sum_{t \in \mathcal{T}_V} \sum_{a \in \mathcal{C}} w_{ci}^{a} \cdot h(I, t; \triangle t_C) \mathbb{I}(a, C(t)) \tag{2}$$

where w_i is the vector of model weights for each class of collective activity, $h(I, t; \triangle t_C)$ is an $\mathcal{I}$ dimensional histogram function of interaction labels around time t (within a temporal window $\pm\triangle t_C$), and $\mathbb{I}(\cdot, \cdot)$ is an indicator function, that returns 1 if the two inputs are the same and 0 otherwise.

Collective Activity Transition $\Psi(C)$: This potential models the temporal smoothness of collective activities across adjacent frames. That is,

$$\Psi(C) = \sum_{t \in \mathcal{T}_V} \sum_{a \in \mathcal{C}} \sum_{b \in \mathcal{C}} w_{c}^{ab} \; \mathbb{I}(a, C(t)) \; \mathbb{I}(b, C(t+1)) \tag{3}$$

Interaction Transition $\Psi(I) = \sum_{i,j} \Psi(I_{ij})$: This potential models the temporal smoothness of interactions across adjacent frames. That is,

$$\Psi(I_{ij}) = \sum_{t \in \mathcal{T}_V} \sum_{a \in \mathcal{I}} \sum_{b \in \mathcal{I}} w_{i}^{ab} \; \mathbb{I}(a, I_{ij}(t)) \; \mathbb{I}(b, I_{ij}(t+1)) \tag{4}$$

Interaction - Atomic $\Psi(I, A, T) = \sum_{i,j} \Psi(A_i, A_j, I_{ij}, T)$: This encodes the correlation between the interaction I_{ij} and the relative motion between two atomic motions A_i and A_j given all target associations T (more precisely the trajectories of T_k and T_l to which τ_i and τ_j belong, respectively). The relative motion is

encoded by the feature vector ψ and the potential $\Psi(A_i, A_j, I_{ij}, T)$ is modelled as:

$$\Psi(A_i, A_j, I_{ij}, T) = \sum_{t \in \mathcal{T}_V} \sum_{a \in \mathcal{I}} w_{ai}^a \cdot \psi(A_i, A_j, T, t; \triangle t_I)\ \mathbb{I}(a, I_{ij}) \quad (5)$$

where $\psi(A_i, A_j, T, t; \triangle t_I)$ is a vector representing the relative motion between two targets within a temporal window $(t - \triangle t_I,\ t + \triangle t_I)$ and w_{ai}^a is the model parameter for each class of interaction. The feature vector is designed to encode the relationships between the locations, poses, and actions of two people. See [29] for details. Note that since this potential incorporates information about the location of each target, it is closely related to the problem of target association. The same potential is used in both the activity classification and the multi-target tracking components of our framework.

Atomic Prior $\Psi(A)$: Assuming independence between pose and action, the function is modelled as a linear sum of pose transition $\Psi_p(A)$ and action transition $\Psi_a(A)$. This potential function is composed of two functions that encode the smoothness of pose and action. Each of them is parameterized as the co-occurrence frequency of the pair of variables similar to $\Psi(I_{ij})$.

Observations $\Psi(A, O) = \sum_i \Psi(A_i, O_i)$ and $\Psi(C, O)$: these model the compatibility of atomic (A) and collective (C) activity with observations (O). Details of the features are explained in Sec.7.

4 Multiple Target Tracking

Our multi-target tracking formulation follows the philosophy of [30], where tracks are obtained by associating corresponding tracklets. Unlike other methods, we leverage the contextual information provided by interaction activities to make target association more robust. Here, we assume that a set of initial tracklets, atomic activities, and interaction activities are given. We will discuss the joint estimation of these labels in Sec.5.

As shown in Fig.3, tracklet association can be formulated as a min-cost network problem [15], where the edge between a pair of nodes represents a tracklet, and the black directed edges represent possible links to match two tracklets. We refer the reader to [15, 16] for the details of network-flow formulations.

Given a set of tracklets $\tau_1, \tau_2, ..., \tau_N$ where $\tau_i = \{x_{\tau_i}(t_0^i), ..., x_{\tau_i}(t_e^i)\}$ and $x(t)$ is a position at t, the tracklet association problem can be stated as that of finding an unknown number M of associations $T_1, T_2, ..., T_M$, where each T_i contains one or more indices of tracklets. For example, one association may consist of tracklets 1 and 3: $T_1 = \{1, 3\}$. To accomplish this, we find a set of possible paths between two non-overlapping tracklets τ_i and τ_j. These correspond to match hypotheses $p_{ij}^k = \{x_{p_{ij}^k}(t_e^i + 1), ..., x_{p_{ij}^k}(t_0^j - 1)\}$ where the timestamps are in the temporal gap between τ_i and τ_j. The association T_i can be redefined by augmenting the associated pair of tracklets τ_i and τ_j with the match hypothesis p_{ij}. For example, $T_1 = \{1, 3, 1\text{-}2\text{-}3\}$ indicates that tracklet 1 and 3 form one track and the second match hypothesis (the solid edge between τ_1 and τ_3 in Fig. 3) connects them. Given human detections, we can generate match hypotheses using the K-shortest path algorithm [31] (see [29] for details).

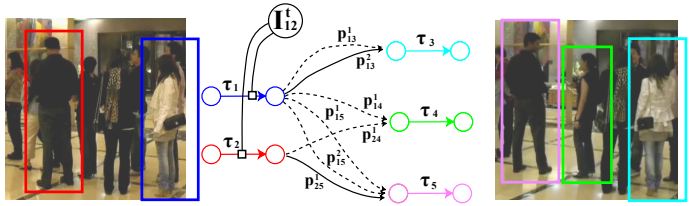


Fig. 3: The tracklet association problem is formulated as a min-cost flow network [15, 16]. The network graph is composed of two components: tracklets τ and path proposals p. In addition to these two, we incorporate interaction potential to add robustness in tracklet association. In this example, the interaction "standing-in-a-row" helps reinforce the association between tracklets τ_1 and τ_3 and penalizes the association between τ_1 and τ_4.

Each match hypothesis has an associated cost value c_{ij}^k that represents the validity of the match. This cost is derived from detection responses, motion cues, and color similarity. By limiting the number of hypotheses to a relatively small value of K, we prune out a majority of the exponentially many hypotheses that could be generated by raw detections. If we define the cost of entering and exiting a tracklet as c_{en} and c_{ex} respectively, the tracklet association problem can be written as :

$$\hat{f} = \underset{f}{\operatorname{argmin}}\, c^T f = \underset{f}{\operatorname{argmin}} \sum_i c_{en} f_{en,i} + \sum_i c_{ex} f_{i,ex} + \sum_{i,j} \sum_k c_{ij}^k f_{ij}^k$$

$$s.t.\ f_{en,i}, f_{i,ex}, f_{ij}^k \in \{0,1\},\ f_{en,i} + \sum_j \sum_k f_{ji}^k = f_{i,ex} + \sum_j \sum_k f_{ij}^k = 1$$

where f represent the flow variables, the first set of constraints is a set of binary constraints and the second one captures the inflow-outflow constraints (we assume all the tracklets are true). Later in this paper, we will refer to $\mathbb{S}$ as the feasible set for f that satisfies the above constraints. Once the flow variable f is specified, it is trivial to obtain the tracklet association T through a mapping function $T(f)$. The above problem can be efficiently solved by binary integer programming, since it involves only a few variables, with complexity $O(KN)$ where N (the number of tracklets) is typically a few hundred, and there are $2N$ equality constraints. Note that the number of nodes in [15, 16] is usually in the order of tens or hundreds of thousands.

One of the novelties of our framework lies in the contextual information that comes from the interaction activity nodes. For the moment, assume that the interactions I_{12}^t between A_1 and A_2 are known. Then, selecting a match hypothesis f_{ij}^k should be related with the likelihood of observing the interaction I_{12}^t. For instance, the *red* and *blue* targets in Fig.3 are engaged in the *standing-in-a-row* interaction activity. If we select the match hypothesis that links *red* with *pink* and *blue* with *sky-blue* (shown with solid edges), then the interaction will be compatible with the links, since the distance between *red* and *blue* is similar to that between *pink*/*sky-blue*. However, if we select the match hypothesis that links *red* with *green*, this will be less compatible with the *standing-in-a-row* interaction activity, because the *green*/*pink* distance is less than the *red*/*blue* distance, and people do not tend to move toward each other when they are in a queue. The potential $\Psi(I, A, T)$ (Sec.3.2) is used to enforce this consistency between interactions and tracklet associations.

5 Unifying activity classification and tracklet association

The previous two sections present collective activity classification and multi-target tracking as independent problems. In this section, we show how they can be modelled in a unified framework. Let $\hat{y}$ denote the desired solution of our unified problem. The optimization can be written as:

$$\hat{y} = \underset{f,C,I,A}{\operatorname{argmax}} \underbrace{\Psi(C, I, A, O, T(f))}_{Sec.3} - \underbrace{c^T f}_{Sec.4}, \ s.t. \ f \in \mathbb{S} \tag{6}$$

where f is the binary flow variables, $\mathbb{S}$ is the feasible set of f, and C, I, A are activity variables. As noted in the previous section, the interaction potential $\Psi(A, I, T)$ involves the variables related to both activity classification (A, I) and tracklet association (T). Thus, changing the configuration of interaction and atomic variables affects not only the energy of the classification problem, but also the energy of the association problem. In other words, our model is capable of propagating the information obtained from collective activity classification to target association and from target association to collective activity classification through $\Psi(A, I, T)$.

5.1 Inference

Since the interaction labels I and the atomic activity labels A guide the flow of information between target association and activity classification, we leverage the structure of our model to efficiently solve this complicated joint inference problem. The optimization problem Eq.6 is divided into two sub problems and solved iteratively:

$$\{\hat{C}, \hat{I}, \hat{A}\} = \underset{C,I,A}{\operatorname{argmax}} \Psi(C, I, A, O, T(\hat{f})) \ AND \ \hat{f} = \underset{f}{\operatorname{argmin}} c^T f - \Psi(\hat{I}, \hat{A}, T(f)), \ s.t. \ f \in \mathbb{S} \tag{7}$$

Given $\hat{f}$ (and thus $\hat{T}$) the hierarchical classification problem is solved by applying iterative Belief Propagation. Fixing the activity labels A and I, we solve the target association problem by applying the Branch-and-Bound algorithm with a tight linear lower bound (see below for more details).

Iterative Belief Propagation. Due to the high order potentials in our model (such as the Collective-Interaction potential), the exact inference of the all variables is intractable. Thus, we propose an approximate inference algorithm that takes advantage of the structure of our model. Since each type of variable forms a simple chain in the temporal direction (see Fig.1), it is possible to obtain the optimal solution given all the other variables by using belief propagation [32].

Algorithm 1 Iterative Belief Propagation

Require: Given association $\hat{T}$ and observation O.
  Initialize C^0, I^0, A^0
  while Convergence, k++ **do**
    $C^k \Leftarrow \operatorname{argmax}_C \Psi(C, I^{k-1}, A^{k-1}, O, \hat{T})$
    for all $\forall i \in A$ **do**
      $A_i^k \Leftarrow \operatorname{argmax}_A \Psi(C^k, I^{k-1}, A, A_{\setminus i}^{k-1}, O, \hat{T})$
    end for
    for all $\forall i \in I$ **do**
      $I_i^k \Leftarrow \operatorname{argmax}_I \Psi(C^k, I, I_{\setminus i}^{k-1}, A^k, O, \hat{T})$
    end for
  end while

The iterative belief propagation algorithm is grounded in this intuition, and is shown in detail in Alg.1.

Target Association Algorithm. We solve the association problem by using the Branch-and-Bound method. Unlike the original min-cost flow network problem, the interaction terms introduce a quadratic relationship between flow variables. Note that we need to choose at most two flow variables to specify one interaction feature. For instance, if there exist two different tails of tracklets at the same time stamp, we need to specify two of the flows out of seven flows to compute the interaction potential as shown in Fig.3. This leads to a non-convex binary quadratic programming problem which is hard to solve exactly (the Hessian H is not a positive semi-definite matrix).

$$\underset{f}{\operatorname{argmin}} \frac{1}{2} f^T H f + c^T f, \ s.t. \ f \in \mathbb{S} \tag{8}$$

To tackle this issue, we use a Branch-and-Bound (BB) algorithm with a novel tight lower bound function given by $h^T f \leq \frac{1}{2} f^T H f, \ \forall f \in \mathbb{S}$. See [29] for details about variable selection, lower and upper bounds, and definitions of the BB algorithm.

6 Model Learning

Given the training videos, the model is learned in a two-stage process: i) learning the observation potentials $\Psi(A, O)$ and $\Psi(C, O)$. This is done by learning each observation potential $\Psi(\cdot)$ independently using multiclass SVM [28]. ii) learning the model weights w for the full model in a max-margin framework as follows. Given a set of N training videos (x^n, y^n), $n = 1, ..., N$, where x^n is the observations from each video and y^n is a set of labels, we train the global weight w in a max-margin framework. Specifically, we employ the cutting plane training algorithm described in [33] to solve this optimization problem. We incorporate the inference algorithm described in Sec.5.1 to obtain the most violated constraint in each iteration [33]. To improve computational efficiency, we train the model weights related to activity potentials first, and train the model weights related to tracklet association using the learnt activity models.

7 Experimental Validation

Implementation details. Our algorithm assumes that the inputs O are available. These inputs are composed of collective activity features, tracklets, appearance feature, and spatio-temporal features as discussed in Sec.3.1. Given a video, we obtain tracklets using a proper tracking method (see text below for details). Once tracklets O are obtained, we compute two visual features (the histogram of oriented gradients (HoG) decriptors [26] and the bag of video words (BoV) histogram [17]) in order to classify poses and actions, respectively. The HoG is extracted from an image region within the bounding box of the tracklets and the BoV is constructed by computing the histogram of video-words within the spatio-temporal volume of each tracklet. To obtain the video-words, we apply PCA (with 200 dimensions) and the k-means algorithm (100 codewords) on the cuboids obtained by [17]. Finally, the collective activity features are computed using the STL descriptor [1] on tracklets and pose classification estimates. We

	Dataset [1]				New Dataset			
Method	Ovral (C)	Mean (C)	Ovral (I)	Mean (I)	Ovral (C)	Mean (C)	Ovral (I)	Mean (I)
without O_C	38.7	37.1	40.5	37.3	59.2	57.4	49.4	41.1
no edges between C and I	67.7	68.2	42.8	37.7	67.8	54.6	42.4	32.8
no temporal chain	66.9	66.3	42.6	33.7	71.1	68.9	41.9	46.1
no temporal chain between C	74.1	75.0	54.2	48.6	77.0	76.1	**55.9**	**48.6**
full model ($\triangle t_C = 20, \triangle t_I = 25$)	**79.0**	**79.6**	**56.2**	**50.8**	**83.0**	**79.2**	53.3	43.7
baseline	72.5	73.3	-	-	77.4	74.3	-	-

Table 1: Comparison of collective and interaction activity classification for different versions of our model using the dataset [1] (left column) and the newly proposed dataset (right column). The models we compare here are: i) *Graph without* O_C. We remove observations (STL [1]) for the collective activity. ii) *Graph with no edges between* C *and* I. We cut the connections between variables C and I and produce separate chain structures for each set of variables. iii) *Graph with no temporal edges.* We cut all the temporal edges between variables in the graphical structure and leave only hierarchical relationships. iv) *Graph with no temporal chain between* C *variables.* v) Our full model shown in Fig.1.(d) and vi) baseline method. The baseline method is obtained by taking the max response from the collective activity observation (O_C).

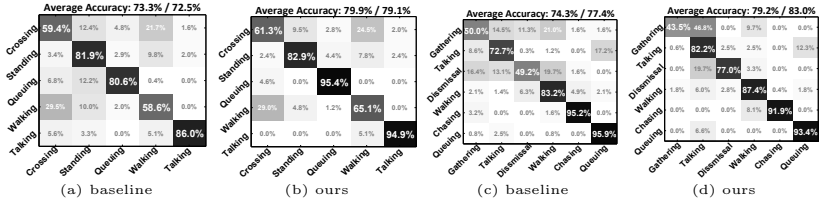
	Dataset [1]				New Dataset			
Method	Ovral (C)	Mean (C)	Ovral (I)	Mean (I)	Ovral (C)	Mean (C)	Ovral (I)	Mean (I)
$\triangle t_C = 30, \triangle t_I = 25$	79.1	79.9	56.1	50.8	80.8	77.0	**54.3**	**46.3**
$\triangle t_C = 20, \triangle t_I = 25$	79.0	79.6	**56.2**	**50.8**	**83.0**	**79.2**	53.3	43.7
$\triangle t_C = 10, \triangle t_I = 25$	77.4	78.2	56.1	50.7	81.5	77.6	52.9	41.8
$\triangle t_C = 30, \triangle t_I = 15$	76.1	76.7	52.8	40.7	80.7	71.8	48.6	34.8
$\triangle t_C = 30, \triangle t_I = 5$	**79.4**	**80.2**	45.5	36.6	77.0	67.3	37.7	25.7

Table 2: Comparison of classification results using different lengths of temporal support $\triangle t_C$ and $\triangle t_I$ for collective and interaction activities, respectively. Notice that in general larger support provides more stable results.

adopt the parameters suggested by [1] for STL construction (8 meters for maximum radius and 60 frames for the temporal support). Since we are interested in labelling one collective activity per one time slice (i.e. a set of adjacent time frames), we take the average of all collected STL in the same time slice to generate an observation for C. In addition, we append the mean of the HoG descriptors obtained from all people in the scene to encode the shape of people in a certain activity. Instead of directly using raw features from HoG, BoV, and STL, we train multiclass SVM classifiers [33] for each of the observations to keep the size of parameters within a reasonable bound. In the end, each of the observation features is represented as a $|\mathcal{P}|$, $|\mathcal{A}|$, and $|\mathcal{C}|$ dimensional features, where each dimension of the features is the classification score given by the SVM classifier. In the experiments, we use the SVM response for C as a baseline method (Tab.1 and Fig.4).

Given tracklets and associated pose/action features O, a temporal sequence of atomic activity variables A_i is assigned to each tracklet τ_i. For each pair of coexisting A_i and A_j, I_{ij} describes the interaction between the two. Since I is defined over a certain temporal support ($\triangle t_I$), we sub-sample every 10th frames to assign an interaction variable. Finally, one C variable is assigned in every 20 frames with a temporal support $\triangle t_C$. We present experimental results using different choices of $\triangle t_I$ and $\triangle t_C$, (Tab.2). Given tracklets and observations (O and O_C), the classification and target association take about a minute per video in our experiments.

Datasets and experimental setup. We present experimental results on the public dataset [1] and a newly proposed dataset. The first dataset is composed of 44 video clips with annotations for 5 collective activities (*crossing*, *waiting*, *queuing*, *walking*, and *talking*) and 8 poses (*right*, *right-front*, ..., *right-back*). In addition to these labels, we annotate the target correspondence, action labels and interaction labels for all sequences. We define the 8 types of interactions as *approaching* (AP), *leaving* (LV), *passing-by* (PB), *facing-each-other* (FE),



(a) baseline (b) ours (c) baseline (d) ours

Fig. 4: (a) and (b) shows the confusion table for collective activity using baseline method (SVM response for C) and proposed method on dataset [1], respectively. (c) and (d) compare the two methods on newly proposed dataset. In both cases, our full model improves the accuracy significantly over the baseline method. The numbers on top of each table show *mean-per-class* and *overall* accuracies.

walking-side-by-side (WS), *standing-in-a-row* (SR), *standing-side-by-side* (SS) and *no-interaction* (NA). The categories of atomic actions are defined as: *standing* and *walking*. Due to a lack of standard experimental protocol on this dataset, we adopt two experimental scenarios. First, we divide the whole set into 4 subsets without overlap of videos and perform 4-fold training and testing. Second, we divide the set into separate training and testing sets as suggested by [11]. Since the first setup provides more data to be analysed, we run the main analysis with the setup and use the second for comparison against [11]. In the experiments, we use the tracklets provided on the website of the authors of [6, 1].

The second dataset is composed of 32 video clips with 6 collective activities: *gathering*, *talking*, *dismissal*, *walking together*, *chasing*, *queueing*. For this dataset, we define 9 interaction labels: *approaching* (AP), *walking-in-opposite-direction* (WO), *facing-each-other* (FE), *standing-in-a-row* (SR), *walking-side-by-side* (WS), *walking-one-after-the-other* (WR), *running-side-by-side* (RS), *runn ing-one-after-the-other* (RR), and *no-interaction* (NA). The atomic actions are labelled as *walking*, *standing still*, and *running*. We define 8 poses similarly to the first dataset. We divide the whole set into 3 subsets and run 3-fold training and testing. For this dataset, we obtain the tracklets using [16] and create back projected 3D trajectories using the simplified camera model [34].

Results and Analysis. We analyze the behavior of the proposed model by disabling the connectivity between various variables of the graphical structure (see Tab.1 and Fig.4 for details). We study the classification accuracy of collective activities C and interaction activities I. As seen in the Tab.1, the best classification results are obtained by our full model. Since the dataset is unbalanced, we present both overall accuracy and mean-per-class accuracy, denoted as Ovral and Mean in Tab.1 and Tab.2.

Next, we analyse the model by varying the parameter values that define the temporal supports of collective and interaction activities ($\triangle t_C$ and $\triangle t_I$). We run different experiments by fixing one of the temporal supports to a reference value and change the other. As any of the temporal supports becomes larger, the collective and interaction activity variables are connected with a larger number of interactions and atomic activity variables, respectively, which provides richer coupling between variables across labels of the hierarchy and, in turn, enables more robust classification results (Tab.2). Notice that, however, by increasing connectivity, the graphical structure becomes more complex and thus inference becomes less manageable.

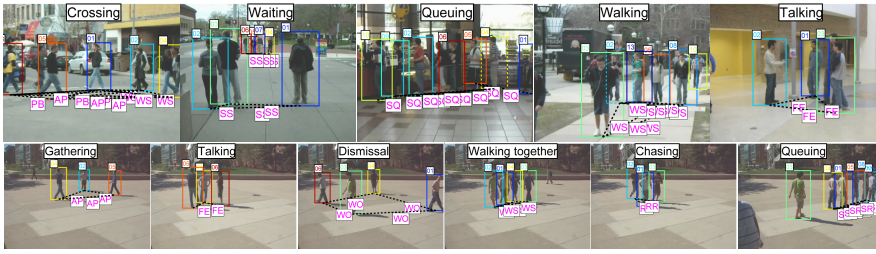


Fig. 5: Anecdotal results on different types of collective activities. In each image, we show the collective activity estimated by our method. Interactions between people are denoted by the dotted line that connects each pair of people. To make the visualization more clear, we only show interactions that are not labelled as NA (*no interaction*). Anecdotal results on the dataset [1] and the newly proposed dataset are shown on the top and bottom rows, respectively. Our method automatically discovers the interactions occurring within each collective activity; Eg. *walking-side-by-side* (denoted as WS) occurs with *crossing* or *walking*, whereas *standing-side-by-side* (SS) occurs with *waiting*. See text for the definition of other acronyms.

Since previous works adopt different ways of calculating the accuracy of the collective activity classification, a direct comparison of the results may not be appropriate. [1] and [2] adopt leave-one-video-out training/testing and evaluate per-person collective activity classification. [11] train their model on three fourths of the dataset, test on the remaining fourth and evaluate per-scene collective activity classification. To compare against [1, 2], we assign the per-scene collective activity labels that we obtain with four-fold experiments to each individual. We obtain an accuracy of 74.4% which is superior than 65.9% and 70.9% reported in [1] and [2], respectively. In addition, we run the experiments on the same training/testing split of the dataset suggested by [11] and achieve competitive accuracy (80.4% overall and 75.7% mean-per-class compared to 79.1% overall and 77.5% mean-per-class, respectively, reported in [11]). Anecdotal results are shown in Fig.5.

Tab.3 summarizes the tracklet association accuracy of our method. In this experiment, we test three different algorithms for tracklet matching : pure match, linear model, and full quadratic model. *Match* represents the max-flow method without interaction potential (only appearance, motion and detection scores are used). *Linear* model represents our model where the quadratic relationship is ignored and only the linear part of the interaction potentials is considered (e.g. those interactions that are involved in selecting only one path). The *Quadratic* model represents our full Branch-and-Bound method for target association. The estimated activity labels are assigned to each variable for the two methods. We also show the accuracy of association when ground truth (GT) activity labels are provided, in the fourth and fifth columns of the table. The last column shows the number of association errors in the initial input tracklets. In these experiments, we adopt the same four fold training/testing and three fold training/testing for the dataset [1] and newly proposed dataset, respectively. Note that, in the dataset [1], there exist 1821 tracklets with 1556 match errors in total. In the new dataset, which includes much less crowded sequences than [1], there exist 474 tracklets with 604 errors in total. As the Tab.3 shows, we achieve significant improvement over baseline method (*Match*) using the dataset [1] as it is more challenging and involves a large number of people (more information from interactions). On the other hand, we observe a smaller improvement in matching

(a) (b)

Fig. 6: The discovered interaction *standing-side-by-side* (denoted as SS) helps to keep the identity of tracked individuals after an occlusion. Notice the complexity of the association problem in this example. Due to the proximity of the targets and similarity in color, the *Match* method **(b)** fails to keep the identity of targets. However, our method **(a)** finds the correct match despite the challenges. The input tracklets are shown as a solid box and associated paths are shown in dotted box.

	Match (baseline)	*Linear* (partial model)	*Quadratic* (full model)	*Linear GT*	*Quad. GT*	*Tracklet*
Dataset [1]	1109/28.73%	974/37.40%	894/42.54%	870/44.09%	736/52.70%	1556/0%
New Dataset	110/81.79%	107/82.28%	104/82.78%	97/83.94%	95/84.27%	604/0%

Table 3: Quantitative tracking results and comparison with baseline methods (see text for definitions). Each cell of the table shows the number of match errors and Match Error Correction Rate (MECR) $\frac{\#\ error\ in\ tracklet\ -\ \#\ error\ in\ result}{\#\ error\ in\ tracklet}$ of each method, respectively. Since we focus on correctly associating each tracklet with another, we evaluate the method by counting the number of errors made during association (rather than detection-based accuracy measurements such as recall, FPPI, etc) and MECR. An association error is defined for each possible match of a tracklet (thus at most two per tracklets, previous and next match). This measure can effectively capture the amount of fragmentization and identity switches in association. In the case of a false alarm tracklet, any association with this track is considered to be an error.

targets in the second dataset, since it involves few people (typically $2 \sim 3$) and is less challenging (note that the baseline (*Match*) already achieves 81% correct match). Experimental results obtained with ground truth activity labels (*Linear GT* and *Quad. GT*) suggest that better activity recognition would yield more accurate tracklet association. Anecdotal results are shown in Fig.6.

8 Conclusion

In this paper, we present a new framework to coherently identify target associations and classify collective activities. We demonstrate that collective activities provide critical contextual cues for making target association more robust and stable; in turn, the estimated trajectories as well as atomic activity labels allow the construction of more accurate interaction and collective activity models.

Acknowledgement: We acknowledge the support of the ONR grant N00014111 0389 and Toyota. We appreciate Yu Xiang for his valuable discussions.

References

1. Choi, W., Shahid, K., Savarese, S.: What are they doing? : Collective activity classification using spatio-temporal relationship among people. In: VSWS. (2009)
2. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR. (2011)
3. Scovanner, P., Tappen, M.: Learning pedestrian dynamics from the real world. In: ICCV. (2009)
4. Pellegrini, S., Ess, A., Schindler, K., van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV. (2009)
5. Leal-Taixe, L., Pons-Moll, G., Rosenhahn, B.: Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In: Workshop on Modeling, Simulation and Visual Analysis of Large Crowds, ICCV. (2011)
6. Choi, W., Savarese, S.: Multiple target tracking in world coordinate with single, minimally calibrated camera. In: ECCV. (2010)
7. Khan, Z., Balch, T., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. PAMI (2005)

8. Yamaguchi, K., Berg, A.C., Berg, T., Ortiz, L.: Who are you with and where are you going? In: CVPR. (2011)
9. Intille, S., Bobick, A.: Recognizing planned, multiperson action. CVIU (2001)
10. Li, R., Chellappa, R., Zhou, S.K.: Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition. In: CVPR. (2009)
11. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: NIPS. (2010)
12. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. IJCV (2007)
13. Ess, A., Leibe, B., Schindler, K., , van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR. (2008)
14. Rodriguez, M., Ali, S., Kanade, T.: Tracking in unstructured crowded scenes. In: ICCV. (2009)
15. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR. (2008)
16. Pirsiavash, H., Ramanan, D., Fowlkes, C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR. (2011)
17. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS. (2005)
18. Savarese, S., DelPozo, A., Niebles, J., Fei-Fei, L.: Spatial-temporal correlatons for unsupervised action classification. In: WMVC. (2008)
19. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. IJCV (2008)
20. Liu, J., Luo, J., Shah, M.: Recongizing realistic actions from videos "in the wild". In: CVPR. (2009)
21. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV. (2009)
22. Swears, E., Hoogs, A.: Learning and recognizing complex multi-agent activities with applications to american football plays. In: WACV. (2011)
23. Ni, B., Yan, S., Kassim, A.: Recognizing human group activities with localized causalities. In: CVPR. (2009)
24. Ryoo, M.S., Aggarwal, J.K.: Stochastic representation and recognition of high-level group activities. IJCV (2010)
25. Ramin Mehran, A.O., Shah, M.: Abnormal crowd behavior detection using social force model. In: CVPR. (2009)
26. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
27. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. MIT Press (2006)
28. Weston, J., Watkins, C.: Multi-class support vector machines (1998)
29. Choi, W., Savarese, S.: Supplementary material. In: ECCV. (2012)
30. Singh, V.K., Wu, B., Nevatia, R.: Pedestrian tracking by associating tracklets using detection residuals. In: IMVC. (2008)
31. Yen, J.Y.: Finding the k shortest loopless paths in a network. (Management Science)
32. Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. In: IJCV. (2006)
33. Joachims, T., Finley, T., Yu, C.N.: Cutting-plane training of structural svms. Machine Learning (2009)
34. Hoiem, D., Efros, A.A., Herbert, M.: Putting objects in perspective. IJCV (2008)