Learning Context for Collective Activity Recognition

Wongun Choi
wgchoi@umich.edu

Khuram Shahid
kshahid@umich.edu

Silvio Savarese
silvio@eecs.umich.edu

Dept. of Electrical and Computer Engineering, University of Michigan
1301 Beal Avenue, Ann Arbor, MI 48109-2122

Abstract

In this paper we present a framework for the recognition of collective human activities. A collective activity is defined or reinforced by the existence of coherent behavior of individuals in time and space. We call such coherent behavior 'Crowd Context'. Examples of collective activities are "queuing in a line" or "talking". Following [7], we propose to recognize collective activities using the crowd context and introduce a new scheme for learning it automatically. Our scheme is constructed upon a Random Forest structure which randomly samples variable volume spatio-temporal regions to pick the most discriminating attributes for classification. Unlike previous approaches, our algorithm automatically finds the optimal configuration of spatio-temporal bins, over which to sample the evidence, by randomization. This enables a methodology for modeling crowd context. We employ a 3D Markov Random Field to regularize the classification and localize collective activities in the scene. We demonstrate the flexibility and scalability of the proposed framework in a number of experiments and show that our method outperforms state-of-the art action classification techniques [7, 19].

1. Introduction

In human interactions, activities have an underlying purpose. This purpose can be to accomplish a goal, or to respond to some stimulus. Both of these parameters are governed by the environment of the individuals, which dictates the contextual elements in the scene. Since this environment is shared by all individuals present in the scene, it is often the case that the actions of individuals are interdependent and some coherency between these actions may exist. We call such activities "collective". Examples of collective activities are: Crossing the road, Talking, Waiting, Queuing, Walking, Dancing and Jogging. In this paper, we seek to recognize such collective activities from videos. We see our work being relevant to a number of applications such as surveillance monitoring, autonomous vehicles, indexing of videos by semantic context and assistive technologies for impaired users in crowded environments.

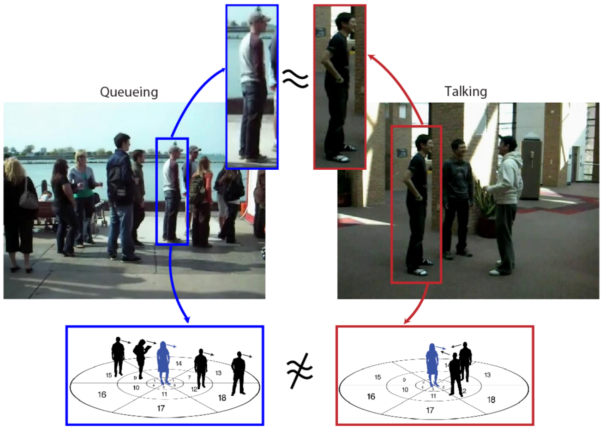


Figure 1. We seek to recognize collective activities such as queuing (left picture) or talking (right picture). In isolation, the highlighted individuals have very similar appearance and thus it is not possible to identify whether they are talking (red) or standing in queue (blue). However, by considering the spatio-temporal distribution of others (i.e, the crowd context) it becomes easier to recognize that the two individuals are performing different activities and to identify which activities are being performed. The spatio-temporal distribution of people, relative to an anchor, is illustrated in the lower part of the figure, where individuals are bucketized over a radial support, based on their physical location in the scene. Such support is discretized in spatio-temporal bins, over which the distribution of individuals can be measured. A key contribution of this paper is to automatically find the optimal configuration of such bins so as to maximize discrimination power.

Consider a collective activity "queuing": the definition of the activity itself requires that multiple individuals be present in the scene and waiting their turn in some structure. Over time, the individuals may progress forward in the queue slowly. This queue structure imposes restrictions on what the spatial distribution of individuals over time may look like. Although the queuing individuals are also "waiting", and a few perhaps are also "talking", the predominant group activity remains the one of queuing. We refer to such dominant and coherent behavior over the temporal and spatial dimension as crowd context. We argue that crowd context is a critical ingredient for characterizing collective activities (Fig.1). Note that if an individual's action differs from those of other individuals in the scene, this individual could be flagged as an outlier since his behavior violates that of the collective activity. For instance, consider a scenario where an individual is in close proximity to a queue but not in the queue. This individual, waiting or talking, instead of queuing would be the anomalous element in the scene.

Characterizing collective activities by crowd context was introduced in [7] and further extended in [19], where a descriptor was introduced to capture coherent behavior around each individual. In this paper we propose to represent crowd context by adaptively binning the spatio-temporal volume as well as the attribute (*e.g.*, pose and velocity of individuals) space using a novel random forest (RF) classification scheme. We call our scheme a Randomized Spatio-Temporal Volume (RSTV) classifier. In our framework, the feature that the trees in a RF operate on, is calculated over a random spatio-temporal volume. Hence, the proposed random forest picks the most discriminating spatio-temporal volume over which to calculate the feature, and then further continues to pick the most discriminating separating plane in order to perform classification as usual in a random forest [6]. Our adaptive binning strategy: 1) establishes robustness to clutter, 2) is able to incorporate other cues/evidence gracefully for classification, and 3) exhibits parameter free learning under a principled probabilistic framework. We use the Random Forest classifier to associate each individual with a collective activity label, performing local classification. We also propose a subsequent step based on a 3D spatio-temporal Markov Random Field that is leveraged to exploit the temporal and spatial consistency of activities to perform global classification.

We point out that the spatial and temporal arrangement solely cannot capture all relevant information for activity classification and it may be well suited to be included as a second layer of abstraction in classification systems such as [28]. Furthermore, it is apparent that collective activity classification can be performed even more accurately given atomic actions (a well studied problem) of each individual. However, in this paper we chose to use only the spatio-temporal distribution of individuals as well as their pose, and no other cues for one reason: to illustrate the vast amount of information implicitly encoded in collective spatial distributions, which we believe is a vital cue that has not been exploited effectively by the vision community. We validate the framework experimentally by comparison to [7, 19] using the dataset of [7].

To summarize, our contributions include: 1) A new representation for capturing crowd context, 2) the use of Random Forest to partition the feature space, 3) the usage of a MRF to regularize collective activities in time and space. This methodology can also be employed to capture the contextual information in other recognition domains such as scene or object-human interaction classification. 4) Validation using a challenging dataset composed of real world and Internet videos.

2. Related Works

Human activity classification has been a key interest in the computer vision community. As a critical ingredient for successful activity classification, researches have developed methods for identifying [13] and tracking [26, 27, 11] humans under different conditions as well as accurately estimating their pose [14, 5]. Activity classification can be regarded as either identification of atomic actions e.g. [16], or recognition of an ensemble of atomic actions that collectively define an activity, such as talking. In this paper we consider the latter. A summary of recent and past literature on action and activity recognition is presented in [29]. Among these, of particular interest are those based on volumetric and contour based representations[3, 30], spatial-temporal filtering[31], distributions of parts [10, 12, 24, 25] sub-volume matching[17] and tensor-based representations [18]. Several of these methods have specific strengths such as modeling self occlusion, being robust to clutter and effectively capturing motion cues. With exception of [16, 21, 32, 23, 26], the focus has remained on recognition of human activities in isolation, independent of the activity of others in the scene. [19] introduced a contextual descriptor which encodes not only the appearance of the person of interest, but the others' appearance information as well. However, the author did not incorporate spatio-temporal relationship among people, which is a critical cue for the collective activity recognition. Inspired by [7], in this paper we consider the crowd context in establishing the activity being performed by each individual in the crowd. Unlike [16, 21], we propose a framework that performs activity recognition by considering the most discriminative cues relevant from the crowd context, which are automatically determined through randomization by a modified Random Forest classifier [6].

3. Framework Overview

This section gives a brief overview of our framework. The essential component of the framework implements the intuition that an individual's behavior is to some extent dictated by the surroundings and can hence be best inferred by considering the relative motion and location of others in the scene (crowd context). To that end, we assume that the trajectories of humans in 3D space as well as human poses are available (Sec.4). We call this the evidence that is used for learning our classification scheme. We then identify the portions of the evidence that are most discriminative in activity classification; this is achieved by randomly sampling hyper-volumes from the evidence space (Sec.5). Finally we perform classification by analyzing the characteristics of the detected individuals over these discriminative regions (Sec.5.3). We demonstrate the ability of our model to successfully categorize collective activities in a number of experiments, as well as to identify individuals in the scene performing incoherent activities (Sec.6).

4. Evidence Extraction

In this section, we explore the methods used to extract the evidence from videos, which is later used for learning and classification. In order to obtain view-point invariant representation of the collective activity, it is desirable to obtain 3D trajectories (x, z) of each individual (see Fig.2). We employed the 3D trajectories produced by [7] which are available online[1]. On top of the provided trajectories, we extract HoG descriptors [9] to obtain appearance information of individuals. In order to classify the pose of individual (left, front, right and back), we incorporate a 4-class linear SVM with the given HoG descriptors.

In order to classify the collective activity of a person of interest (anchor), our algorithm computes the relative motion of others around the anchor. The relative motion includes i) the location (x, z) of others in the coordinate system centered on the anchor and oriented along the anchor's facing direction (pose), ii) the velocity difference between anchor and others, and iii) the pose difference between the anchor and the others (*e.g.,* if the anchor is facing left and an individual is facing right, the relative pose to anchor is defined as facing the opposite direction). These relative motion features contain a number of desired properties to describe a collective activity: 1) invariance under viewpoint change (*e.g.*, camera rotation) 2) consistency within the same category of collective activity. By computing these elements for every person in a certain time window, we can obtain a view-point invariant representation of the contextual distribution of people around the anchor person. This information is passed to the RF classifier in order to obtain the collective activity label for the anchor. Please see next section for more details. Note that unlike [7], our model does not explicitely compute a descriptor for each individual, but the classification algorithm will automatically identify the optimal spatio-temporal structure.

5. The RSTV Model

This section introduces the intuition behind the use of RSTV, its implementation using a Random Forest classifier [Sec. 5.2], as well as the algorithm employed in learning constituent decision trees in the RF and the classification process [Sec. 5.3]. We further outline the motivation behind our system and improvement over state-of-the-art activity classifiers [Sec. 5.1].

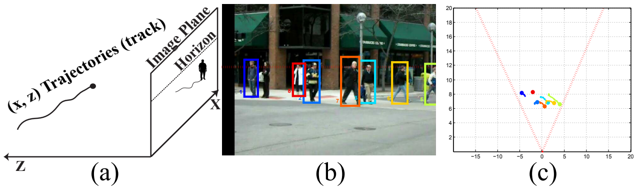
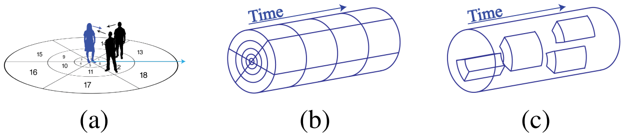


Figure 2. a) Overview of tracking performed in scene coordinates. The algorithm [7] returns a (x,z) trajectory. b) Example view from camera. c) Bird's eye reconstruction of scene in b). Color code is used to indicate different trajectory(tracks) IDs over a time window.

5.1. STV and RSTV

We employ a concept similar to [7], which suggests that successful classification of a collective activity is dependent on considering the spatial distribution of people in the surrounding crowd as well as the temporal evolution of this distribution (crowd context). We start from the intuition that such spatial/temporal distributions can be captured by counting the number of people with a certain pose and velocity in fixed regions of the scene, relative to an anchor person whose activity we would like to classify. We call this representation a Spatio-Temporal Volume representation (STV) (Fig. 3). Our framework extends this intuition and considers variable spatial regions of the scene with a variable temporal support. The full feature space contains all the evidence extracted from the videos: the location of each individual in 3D coordinates as well as the velocity & pose of each individual per video frame. We interpret our representation as a soft binning scheme where the size and locations of bins are estimated by a random forest so as to obtain the most discriminative regions in the feature space. Over these regions, we analyze the density of individuals (Sec.4), which can be used for classification. Figure 3 illustrates the (binned) Spatio-Temporal Volume (STV)[7] and our Randomized Spatio-Temporal Volume (RSTV). RSTV is a generalization of the STV in that the rigid binning restriction imposed in the STV is removed. In the RSTV model, portions of the continuous spatio-temporal volume are sampled at random and the discriminative regions for classification of a certain activity are retained. RSTV provides increasing discrimination power due to increased flexibility (Fig. 3 and Fig. 6). Note that due to this shift toward increasing degrees of freedom, the model complexity and learning time also increase. Nevertheless, testing time does not increase significantly.

There are several reasons behind the proposed framework. 1) As described in Sec.5.3, the framework automatically determines the discriminative features in the whole evidence space for classification. Indeed while STV proposes



(a) (b) (c)

Figure 3. (a) A slice along the cylindrical STV(b) over the temporal axis, which was introduced in [7]. Each slice is associated with a frame from the video. The slice is rotated such that it is oriented in the same direction as the anchor's pose (blue arrow; the anchor is also indicated in blue). The number of people in the scene (in black), with a certain pose and location, both relative to that of the anchor, is histogrammed over several bins. (c) Our proposed RSTV is a generalization of STV and significantly increases the flexibility by allowing variable number of regions as well as variable sized regions. This allows to capture discriminative information about collective activities while also developing robustness to clutter.

a rigid and arbitrary decomposition of the feature space, focused on segmentation over space and time, in RSTV the binning space is partitioned so as to maximize discrimination power. We do so by using our new algorithm based on the random forest decomposition scheme. The algorithm seeks the optimal number, shape and size of the bins given the training set. 2) Unlike [7], there are no parameters that must be learned or selected empirically (e.g. support distance, number of bins). 3) It enables robustness to clutter. Indeed, unlike STV our RSTV does not operate given fixed parameters such as radial support and number of spatial bins, but explores the possible space of all parameters; thus the density feature, using which classification is performed, is only calculated over regions relevant to each different activity. Hence we ensure that the classification evidence is pertinent to each activity and avoid clutter that arises from hard-coded framework parameters that may be tuned to achieve optimal classification of a few activities, but not all. Notice that STV concept is similar to the Shape Context [2] descriptor, which is known to be susceptible to clutter due to non discriminative inclusion of all points within the radial support. See Fig.4 for examples of learnt RSTV regions.

5.2. Random Forest

We propose to use a Random Forest classifier as a key component for learning the structure of RSTV. A Random forest, introduced in [6], is a collection of many singular classifiers known as decision trees. Given a feature, each of these trees is trained to classify the test input by picking a set of decision functions. Bosch *et al.* [4] employ random forests for classification of feature vectors extracted using a spatial pyramid decomposition scheme [20, 15]. However their framework imposes a very rigid binning scheme, specified by empirically determined parameters. Our framework generalizes this pyramiding scheme to allow non-rigid binning which is learned automatically. While [4] provides excellent results because the binning takes place in a 2D feature space (location in the image), a rigid decomposition of our evidence space (x,z,$velocity$,$pose$) would be much less obvious and adequate in our case.

In RSTV, the decision trees are binary and learned top-down. At each branching point in a decision tree, the whole feature space is considered: The algorithm firstly randomizes over different volumes of the feature space and secondly randomizes over different decision thresholds given the feature subspace. In our application, the final classification feature is a scalar count of the number of people that lie within the selected subspace/hyper-volume i.e. $N(S, F, P, V)$. Specifically, presence of a person in a hyper-volume (S, F, P, V) of the feature space indicates that the person is located in some spatial region S (specified by the center position (x, z), radial and angular size ($\triangle r$, $\triangle\theta$)), over a time period of F frames with a pose $P \in (Front, Back, Left, Right)$, with a range of velocity V, where jointly (S, F, P, V) uniquely identify a sub volume. Hence, at each forking node n in the tree, a hyper-volume r_n is selected, the number of people observed within this hyper volume is counted, and upon comparison of this count to a scalar decision threshold, either the left or the right subtree is evaluated. In theory branching on one dimension of the hyper-volume (such as $N(F)$ or $N(V)$) in each node of the tree, as is common, would learn the same information as a tree that branches on the whole feature space: $N(S, F, P, V)$. However, the latter produces a tree of much smaller depth and drastically improves the computational efficiency in testing. At each node of the decision tree, the algorithm picks a discriminating hyper-volume of the feature space and a decision boundary for the feature (number of people) evaluated over the selected hyper-volume. The best hyper-volume and threshold pair can be determined by computing information gain: Eq.1

Regular random forests [6], presented with a fixed dimensional feature vector, pick a decision boundary along one (or multiple) dimension of the vector. By contrast, our feature is in essence of infinite dimensionality since we evaluate the count of people that exist in a certain spatio-temporal region and there exist an uncountably many number of such regions with overlaps. We achieve a computationally feasible implementation through random sampling of the continuous domains.

5.3. Learning RSTV using RF & Classification

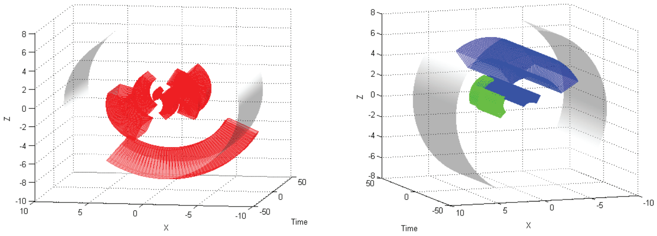
The source of the evidence that is used for classification is the set of tracks T that contain the estimated trajectories, in 3D coordinates, of individuals in the scene. In learning the RF classifier, we begin by growing numerous random decision trees to populate the forest. At each forking node n in the tree we pick a discriminative hyper-volume r_n from the feature space as well as a decision threshold d_n for the count of people present in a set of tracks T evaluated over the selected hyper-volume, denoted by $f(T, r_n)$. This threshold partitions the input into two subsets I_l and I_r. In order to find such a discriminative hyper-volume r_n, the algorithm randomizes over hyper-volumes to automatically determine a sub-optimal solution (Note: sub-optimality is important and is discussed further under Implementaion Considerations). We say a hyper-volume is more discriminating than another if it is better able to separate the training data into 2 sets such that the variance of the activity labels is low within sets and high across sets. This measure is formalized through the notion of information gain ΔE in each trial :

$$\Delta E = -\frac{|I_l|}{|I|}E(I_l) - \frac{|I_r|}{|I|}E(I_r),\ E(I) = -\sum_{i=1}^{C} p_i log_2(p_i) \tag{1}$$

Here I_l and I_r are the partition of set I divided by given feature, C is the number of collective activity classes, p_i is the proportion of collective activity class i in set I, and $|I|$ is the size of the set I. The learning is performed given a learning set I_{train} where I_{train} contains a collection of tracks T of the individuals over some time range with activity labels. Each track contains the location l, velocity v and pose p of individuals present in the scene over the time range for which T is calculated. Hence, T contains $\{l$, v, $p\}$ for all individuals in the scene. Given such an input I, each decision tree should provide the probability that each track in T originates from a person performing some activity - visualizations of learnt RSTV regions are shown in Fig. 4. The procedure for learning r_n and d_n for each node n in a growing tree is outlined below:

```
Learning Algorithm
For each $K^{th}$ tree:
  • Get a random subset $I_k$ of $I$
  • Learn a tree using $I_k$
      – For each node $n$, try $m$ pairs of random ($r_n^m$, $d_n^m$)
      – Select the best ($r_n$, $d_n$) which gives highest information gain.
      – Put training data $\{T|f(T, r_n) < d_n\}$ into $I_l$ and $\{T|f(T, r_n) \geq d_n\}$ into $I_r$.
      – Pass $I_l$ $I_r$ to left and right child node respectively.
      – Recurse until there is only one class of activity.
```

Classification using learned RSTV: For classification, a test set of tracks is passed down each tree in the forest. At each intersection in the tree, the number of people in the learned regions is evaluated and, if this number is larger than the decision threshold, the input is passed to the right subtree, otherwise it is passed to the left subtree. Once a leaf node is reached, the learned posterior for that leaf node is picked up. The posteriors reached in all the decision trees are combined to generate the overall average posterior for class label given the test input.



(a) Waiting (b) Talking

Figure 4. Example of learned RSTV regions. a) & b) each illustrate a set of RSTV regions learned automatically by a single tree. Colors indicate different pose of neighboring individuals. Each RSTV is oriented such that the anchor individual is facing in the z direction. The color red indicates a pose identical to that of the anchor. Hence a) indicates that while waiting, an anchor is surrounded on the left and right by people facing the same direction. The RSTV also captures the presence of individuals standing behind the anchor, who are also facing the same direction (the dataset includes several layers of individuals waiting at a bus stop). RSTV in b) illustrates that during talking the anchor and neighbor face each other and are in very close proximity. The color blue indicates a pose opposite to that of the anchor, while green denotes a side pose. Furthermore the RSTV captures multiple people talking, forming a ring around the anchor. Note that each RSTV needs only capture some coherent portion of evidence since there exist many trees in the RF. x and z have units of meters while time is measured in frames. This figure is best viewed in color

Implementation Considerations: For practical reasons, we consider a maximum radial support of 8 meters as well as a maximum time period of 2 seconds in classification. In learning each tree, while searching for discriminative RSTV regions, it is crucial that the number of random samples of the decision threshold d_n be minimal. We note that although increasing the number of iterations does provide a more optimal r_n, d_n pair, it also decreases the variation among tree nodes and the trees themselves. This is due to the fact that, while learning each node by randomization, we begin to select the same few discriminative r_n, d_n over and over, generating similar trees. It is evident that we would like variation among the trees in order to improve classification by achieving generalization.

5.4. Regularizing Classification using MRF

Though the RSTV captures the local crowd context of a person, the RSTV classification sometimes fails due to noisy estimation of pose or trajectories. To obtain a more robust classifier, we imposed spatial and temporal smoothness by applying 3D Markov Random Field (MRF) [22]. The MRF is depicted in Fig. 5 and formulated as:

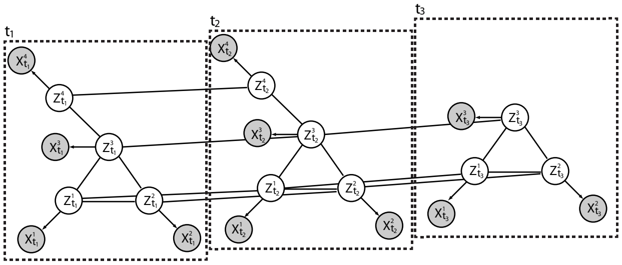


Figure 5. Graphical representation for the proposed MRF. $Z^j_{t_i}$ models the activity of a person in one time slice (hidden variable), $X^j_{t_i}$ represents the trajectories associated to an anchor person. If two people are close enough ($\leq$2 meter away), the spatial edges are built. For every person, temporal edges are constructed between nearby nodes. Linked nodes (either spatially or temporally) defines an activity pair.

$$P(Z|X,p) \propto \prod_t \prod_i P(Z^i_t|X^i_t) \prod_t \prod_{(i,j)\in E_s} \Phi_S(Z^i_t, Z^j_t; p^i_t, p^j_t)$$

$$\prod_i \prod_t \Phi_T(Z^i_{t-1}, Z^i_t) \qquad (2)$$

where Z^i_t is the collective activity label of i^{th} person in time t, X^i_t is the observation of the person in t, p^i_t is the location of person i in t, E_s is the set of edges between people (Fig.5), $P(Z^i_t|X^i_t)$ is the probability estimate from Random Forest for person i in time t, $\Phi_S(Z^i_t, Z^j_t; p^i_t, p^j_t)$ is the spatial pair-wise potential, and $\Phi_T(Z^i_{t-1}, Z^i_t)$ is the temporal potential. We establish temporal edges between temporally adjacent nodes with same target and spatial edges between two nodes belonging to different targets if they are close to each other ($<$ 2 meter). We perform the estimation using Gibbs Sampling (with 500 iterations for burn-in and 1000 iterations for sampling). We learned 1) the temporal potential by counting number of activity pairs in adjacent time and 2) the spatial potential in a non-parametric way by collecting location difference oriented with respect to each person's pose for all activity pairs. In experiments, we use the same leave-one-video-out scheme for validation.

6. Experiments

Here, we provide the results of validating our framework against challenging real world videos.

Dataset: We use the dataset in [7] which is the most appropriate dataset for evaluating collective activities. Other activity datasets (*e.g,* CAVIAR, IXMAS, or UIUC) are not adequate for our purpose, since they either consider only the activity of a single person or few number of people. [7] comprises of two dataset, 5 category collective activity dataset and 6 category collective activity dataset. The former consists of collective activities, *Crossing, Standing, Queuing, Walking* and *Talking*. The later is an augmented dataset based on the former. It includes two more categories of *Dancing* and *Jogging* and omits the activity *Walking*. As suggested in [7], the *Walking* class is ill-defined as it is more like a single person activity than a collective one. We use the trajectory data provided in [1].

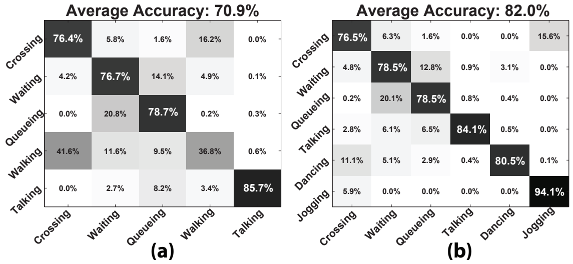


Figure 6. a) Our final classification accuracy for 5-category dataset in [7] is **70.9**%, as compared to 65.9% in [7]. Classification results are obtained using RSTV with MRF regularization. b) Our results for the augmented 6-category dataset in [7].

Activity Classification Results: We validated the performance of our framework against the 5-category and 6-category datasets presented in [7]. Fig.6.(a) and (b) analyzes the performance of our proposed method and shows that we achieve a significant improvement over [7]. We further study the effect of some of the components of our framework and compare it with [19]. Results are shown in Table.1. The first row shows the result of [19] and the second and third row present the result using STV equipped with a SVM classifier [7]. The fourth row shows the result using STV equipped with an RF classifier. This indicates that using STV with a naive replacement of the SVM classifier with a Random Forest does not yield an improvement in the results, indicating that the improvement stems from our novel segmentation of the feature space and selection of relevant portions (RSTV). As Table.1 shows, the MRF yields a much larger improvement if walking is removed (6 activity dataset). All results presented were obtained using a leave-one-video-out scheme. Processing a 1 min video takes roughly 30 min on a standard dual-core desktop. Learning the RSTV takes about 6 min per tree, depending on the size of the learning base. Example results are presented in Fig.9.

Localization Results: Anecdotal results for localization of collective activities and identification of anomalous individual in the scene by MRF are shown in Fig.8. In order to perform the segmentation, we apply mean-shift clustering algorithm [8] over the response of RF classifier.

7. Conclusion

In this paper we demonstrated that capturing crowd context is essential for performing successful classification of collective activities. We presented and validated our

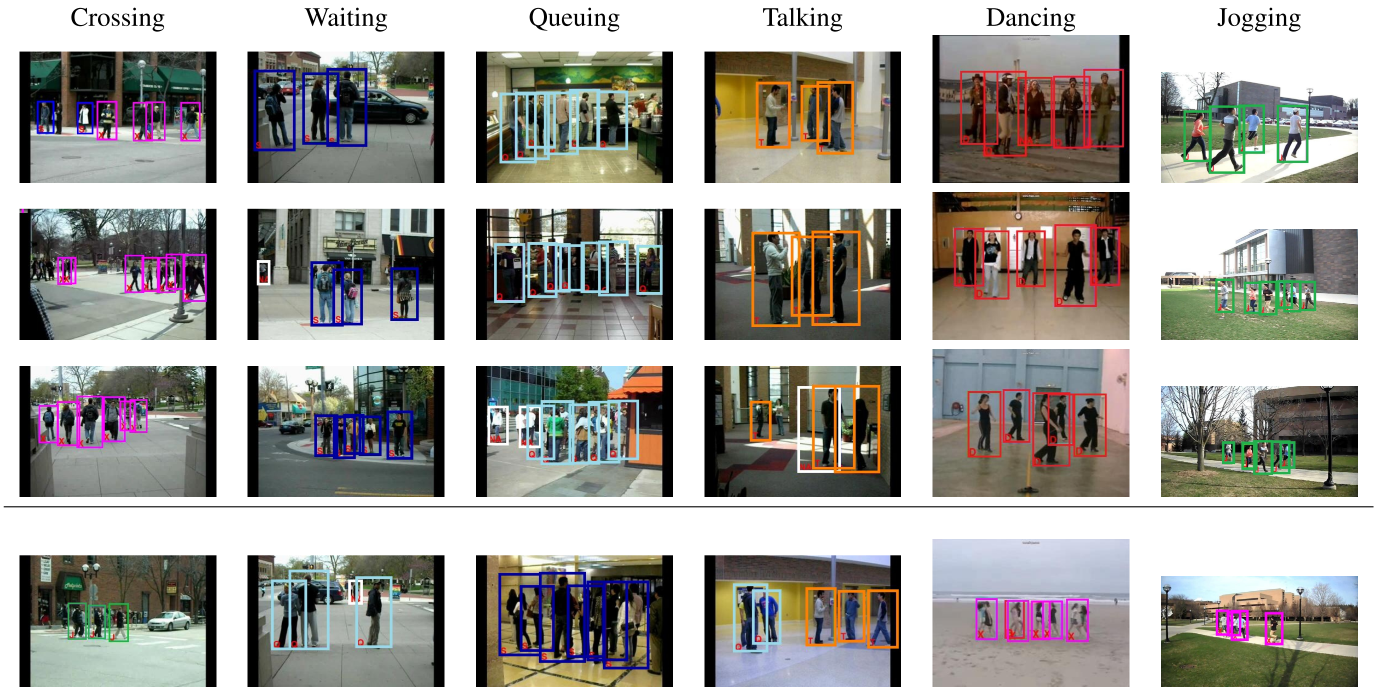


Figure 9. Example results on the 6-category dataset [7]. Top 3 rows show examples of good classification and bottom row shows examples of false classification. The labels X (magenta), S (blue), Q (cyan), T (orange), D (red), J (green) and NA (white) indicate *crossing*, *waiting*, *queuing*, *talking*, *dancing*, *jogging* and not assigned, respectively. When there in insufficient evidence to perform classification, the NA label is displayed. The misclassified results indicate that miss classifications mostly occur between classes with similar structure. This figure is best viewed in color.

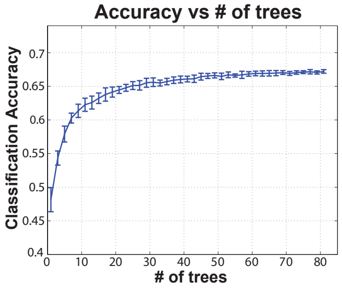


Figure 7. The impact of changing # trees in the forest; 5-category dataset was used, MRF was not employed.

RSTV framework, which is able to automatically capture relevant crowd context and perform activity classification using a random forest. The use of random forest to segment the feature space is a new concept that is of considerable interest as it can be applied to numerous problems in computer vision.

Acknowledgement : This work is supported by a grant from Ford Motor Company via the Ford-U of M Innovation Alliance (Award #N011537) and by NSF AEGER (award #1052762).

Dataset	5 Activities	6 Activities
AC [19]	68.2%	-
STV[7]	64.3%	-
STV+MC[7]	65.9%	-
STV+RF	64.4%	-
RSTV	67.2%	71.7%
RSTV+MRF	**70.9%**	**82.0%**

Table 1. Average classification results for various methods on dataset from [7]. The STV+RF row shows the result of a naive combination of STV with a RF classifier. A comparison with our RSTV results indicates that indeed it is our discriminative learning method that provides the significant improvement. Notice that [7] uses an SVM classifier to classify their STV descriptors.

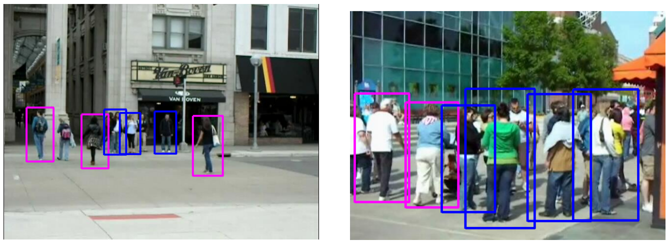
References

[1] Collective activity dataset. `http://www.eecs.umich.edu/vision/activity-dataset.html`. 3275, 3278

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 02. 3276

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, volume 2, Oct. 2005. 3274

[4] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *Proceedings of the*



(a) (b)

Figure 8. Activity Localization. Different collective activities are represented using different color bounding boxes. Activities can be separated using the classification labels: given one predominant activity in the scene, all other activities and individuals that differ from the predominant one are considered anomalous. In (a), one group (crossing) of people are labelled with pink bounding boxes and the other (waiting) are labelled in blue. In (b), queuing people are labelled in blue and walking people are labelled in pink. This figure is best viewed in color.

11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007. 3276

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision*, sep 2009. 3274

[6] L. Breiman and A. Cutler. Random forest. [online], marzec 2004. 3274, 3276

[7] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *Visual Surveillance Workshop, ICCV*, 2009. 3273, 3274, 3275, 3276, 3278, 3279

[8] D. Comaniciu and P. Meer. Mean shift analysis and applications. volume 2, pages 1197–1203 vol.2, 1999. 3278

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, June 05. 3275

[10] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72, Oct. 2005. 3274

[11] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, pages 1–8, June 2008. 3274

[12] C. Fanti, L. Zelnik-Manor, and P. Perona. Hybrid models for human motion recognition. In *CVPR*, volume 1, pages 1166–1173 vol. 1, June 2005. 3274

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, pages 1–8, June 2008. 3274

[14] V. Ferrari, M. Marin-Jiminez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, pages 1–8. IEEE, June 2008. 3274

[15] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. *Computer Vision, IEEE International Conference on*, 2:1458–1465, 2005. 3276

[16] A. Hakeem and M. Shah. Learning, detection and representation of multi-agent events in videos. *Artif. Intell.*, 171(8-9):586–605, 07. 3274

[17] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, volume 1, pages 166–173 Vol. 1, Oct. 2005. 3274

[18] T. Kim, S.-f. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *CVPR*, June 2007. 3274

[19] T. Lan, Y. Wang, G. Mori, and S. Robinovitch. Retrieving actions in group contexts. In *International Workshop on Sign Gesture Activity*, 2010. 3273, 3274, 3278, 3279

[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178, 2006. 3276

[21] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *ECCV*, pages 383–395, Berlin, Heidelberg, 2008. Springer-Verlag. 3274

[22] S. Z. Li. Markov random field models in computer vision. In *ECCV*, 1994. 3277

[23] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *IEEE Conf. Computer Vision and Pattern Recog*, 2009. 3274

[24] J. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, pages 1–8, June 2007. 3274

[25] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, Sep. 2008. 3274

[26] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: modeling social behavior for multi-target tracking. In *IEEE Int'l Conf. on Computer Vision (ICCV)*, 2009. 3274

[27] P. Scovanner and M. F. Tappen. Learning pedestrian dynamics from the real world. *Computer Vision, 2009. ICCV 2009. IEEE 12th International Conference on*, 2009. 3274

[28] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:2004–2011, 2009. 3274

[29] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, Nov. 2008. 3274

[30] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *CVPR*, volume 1, pages 984–989 vol. 1, June 2005. 3274

[31] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, volume 2, pages II–123–II–130 vol.2, 2001. 3274

[32] Y. Zhou, S. Yan, and T. S. Huang. Pair-activity classification by bi-trajectories analysis. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008. 3274