Learning Action Primitives for Multi-Level Video Event Understanding

Tian Lan[1], Lei Chen[2], Zhiwei Deng[2], Guang-Tong Zhou[2], and Greg Mori[2]

Stanford University[1] Simon Fraser University[2]

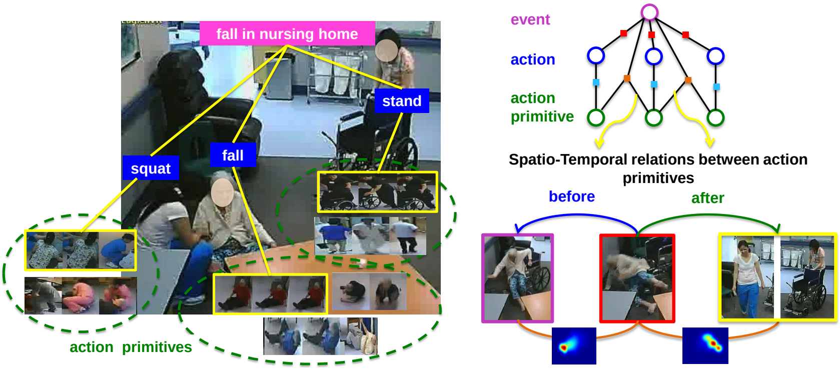
Abstract. Human action categories exhibit significant intra-class variation. Changes in viewpoint, human appearance, and the temporal evolution of an action confound recognition algorithms. In order to address this, we present an approach to discover action primitives, sub-categories of action classes, that allow us to model this intra-class variation. We learn action primitives and their interrelations in a multi-level spatio-temporal model for action recognition. Action primitives are discovered via a data-driven clustering approach that focuses on repeatable, discriminative sub-categories. Higher-level interactions between action primitives and the actions of a set of people present in a scene are learned. Empirical results demonstrate that these action primitives can be effectively localized, and using them to model action classes improves action recognition performance on challenging datasets.

1 Introduction

In recent years, the understanding of complex video events has drawn increased interest in the computer vision community. A complex video event usually involves multiple inter-related people and contains rich spatio-temporal structures at various granularities. Fig. 1 shows an example video event in a long-term care facility. In terms of understanding this type of event, there is a variety of questions one can ask: Is there a fall in the scene? Where is the fallen person? When and how did the person fall? Are there any people coming to help? These potential queries often involve multiple levels of details ranging from the overarching event to the fine-grained details of individuals (*where, when and how*).

In this paper, we develop a novel framework for a complete understanding of video events, including: *event classification* (e.g. fall in nursing home), *action recognition and localization* (e.g. standing, squatting), *fine-grained action primitive discovery* (e.g. pushing a wheelchair, squatting and facing right) and spatio-temporal structure extraction (e.g. squatting beside a person who just fell a few seconds ago).

Understanding complex video events is an extremely challenging problem. It shares all of the difficulties of person detection and action recognition, in addition to significant difficulties unique to event classification. The use of hierarchical models integrating multiple semantics such as actions and/or social roles has been shown to boost event classification performance in realistic videos [1–5]. Despite these successes, there are two important issues not well addressed -



(a) A sample video frame

(b) Our model

Fig. 1. A multi-level video event representation. In the left we show a sample video frame and the goal is to decide whether it is a fall scene or not. We found several actions (e.g. squatting, falling, standing, etc), and each action is recognized using fine-grained action primitives as shown in the yellow rectangles (e.g. squat facing right, stand to push wheelchair, etc). We jointly model the action primitives, actions and event, while considering the spatio-temporal interactions between action primitives. The model for describing the video frame is illustrated in the right side.

localizing the actions and interpreting their fine-grained appearance. The former, usually achieved by generic person detectors (e.g. [6]), brings considerable input noise to the higher-level models, while the latter prevents a deeper level event understanding. We address these problems by modeling *action primitives*, which contain fine-grained information that can not be captured by the basic action categories.

Traditional human detectors are known to have difficulties in capturing the wide range of appearance of human actions in realistic, unconstrained videos. In this paper, we argue that the fine-grained action primitives are key to resolving appearance variations within the action categories. Considering the difficulties in obtaining such labelings, we advocate a weakly supervised setting where the action categories are provided in training, and the action primitives are automatically discovered from the training data. We propose a discriminative spatio-temporal clustering algorithm to discover the action primitives. The action primitives are then treated as mixture components in a latent SVM framework, and refined during learning. Our method detects possibly multiple person instances in each video frame and generates detailed fine-grained action primitives for each instance.

Further, the action primitives naturally contain a rich set of spatial-temporal relations. For example, as shown in Fig. 1, the action primitives: “losing balance”, “lying on the floor” and “pushing wheelchair” are in strict temporal ordering and form typical spatial patterns. These spatio-temporal relations are important to distinguish between different events, such as fall and non-fall. Our model captures these relations, and allows flexible inference of different levels of semantics and their dependencies in a video.

2 Previous Work

The literature on human activity recognition is extensive, and covers a large number of aspects of the problem. A comprehensive review of the field was done by Turaga et al. [7]. In this section we review a selection of closely related research, focusing on spatial and temporal representations and action category learning.

Representations for individuals: A variety of approaches has been developed for representing the action of an individual. Bag-of-words approaches based on local features [8] form the basis for many systems. Recent approaches have pushed toward using a higher-level representation, often by learning mid-level patch representations. Kovashka and Grauman [9] consider higher-order relations between visual words, discriminatively selecting important spatial arrangements. Maji et al. [10] use *poselet* activations, the presence of mid-level body parts indicative of a particular action. Jain et al. [11] learn mid-level discriminative spatio-temporal patches in a data-driven fashion, not relying on poselet-type body part labels.

Many approaches follow a similar vein, analyzing spatio-temporal data to represent human actions in video. Wang et al. [12] track moving points densely over subjects, leading to a dense trajectory feature capturing detailed motion of entire subjects in a scene. Raptis et al. [13] build upon this direction, grouping low-level trajectory-type features into mid-level parts via latent variables. Tian et al. [14] extend the deformable part model to temporal data, modeling the changes in spatio-temporal positions of body parts throughout a sequence. Ma et al. [15] describe a novel multi-scale representation for a person over time, with large and small-scale patches.

Spatio-temporal structures in action recognition: In our work we discover action primitives and model their spatio-temporal relations. Temporal modeling of human actions in terms of lower-level primitives has a long history in computer vision research. The work by Yamato et al. [16] used hidden Markov models (HMMs) and discovered temporal evolution of actions such as tennis swings. Moore and Essa [17] built stochastic grammars to represent components of actions. Bobick and Wilson [18] described state-based representations of gestures. Bregler [19] discovered low-level primitives, again using HMMs, and showed the ability to detect primitives such as states in a gait cycle.

Larger-scale structures, relating the actions of all the individuals in a scene, have been studied previously. Medioni et al. [20] utilized the relative positions and directions of movement of interacting objects such as vehicles and checkpoints. A recent body of work has developed related techniques, trying to infer interactions in video sequences and model arrangements of groups of people. Lan et al. [1, 2] examine latent interaction models at levels of individual actions, social roles, and group activities. Choi et al. [3] unify the problem of inferring activities with tracking individuals. Amer et al. [4] develop a model for multi-scale activity analysis, using AND-OR graphs with efficient inference techniques. Ramanathan et al. [5] learn social roles from weakly labeled data in complex internet videos.

Action localization: Localizing an action in space and time is likely a crucial step in order to reason about group activities. Methods that perform explicit spatio-temporal localization include Ke et al. [21], who develop segmentation-based features for detecting actions in videos with complex, moving background clutter. Klaser et al. [22] track individuals and build features for describing each trajectory before final classification. Lan et al. [23] reason about tracking as a latent variable, and select discriminative sub-regions of a video for classification. Tran and Yuan [24] phrase localization as a regression problem, and learn a structured output model for producing human action bounding boxes in video. As mentioned above, Tian et al. [14] develop deformable part models, which can localize actions spatio-temporally.

Sub-category recognition: A contribution of our work is developing an algorithm for discovering action primitives, sub-categories of the original action classes. In the action recognition literature, this problem has been largely unaddressed. Basic latent variable models have been used, typically modeling aspect or appearance, such as the work of Yao and Fei-Fei [25]. Kitani et al. [26] use a probabilistic latent variable model for discovering action categories. A local feature representation is used, latent action categories are learned over spatial and temporal low-level features. Hoai and Zisserman [27] develop a discriminative approach for sub-category discovery.

In the object recognition community, there exists related work on modeling objects and their subcategories, for instance the work of Lan et al. [28], Gu and Ren [29], and Todorovic and Ahuja [30]. We bridge this line of work to action recognition and develop novel methods for spatio-temporal action sub-category analysis.

3 Action Primitive Based Action Localization

Given a set of training videos with annotations of basic-level action categories and bounding boxes in each frame, our goal is to discover the action primitives and learn action detectors. Our approach is inspired by the recent success of subcategory based object detection [28, 29, 31]. A standard pipeline of this line of work is to first partition the examples of an object category into different subcategories and then learn a multi-component object detector as a mixture of subcategory models. The multi-component object models can handle the intra-class variations and thus improve the object detection performance.

We adapt the multi-component object detectors to the video domain and learn multi-component action detectors. Different from static images, actions in the same video tend to have large correlations, especially when they are temporally close to each other. We propose a novel hierarchical discriminative clustering scheme, to discover action primitives from videos. These action primitives are

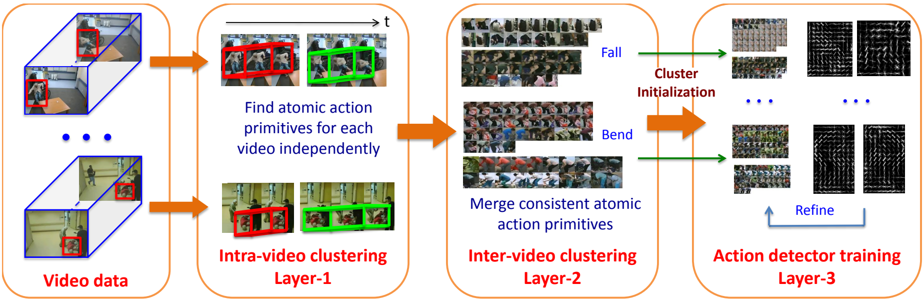

t
Find atomic action primitives for each video independently
Fall
Bend
Cluster Initialization
Merge consistent atomic action primitives
Refine
Video data
Intra-video clustering Layer-1
Inter-video clustering Layer-2
Action detector training Layer-3

Fig. 2. A general pipeline of our action primitive based action localization model. Details are described in the text.

treated as mixture components in the object model, and further refined during learning. An overview of our approach is illustrated in Fig. 2.

Now we introduce the discriminative action clustering algorithm. The algorithm starts by clustering examples in each individual video and then gradually merges consistent clusters from multiple videos. Next, we present each layer in detail.

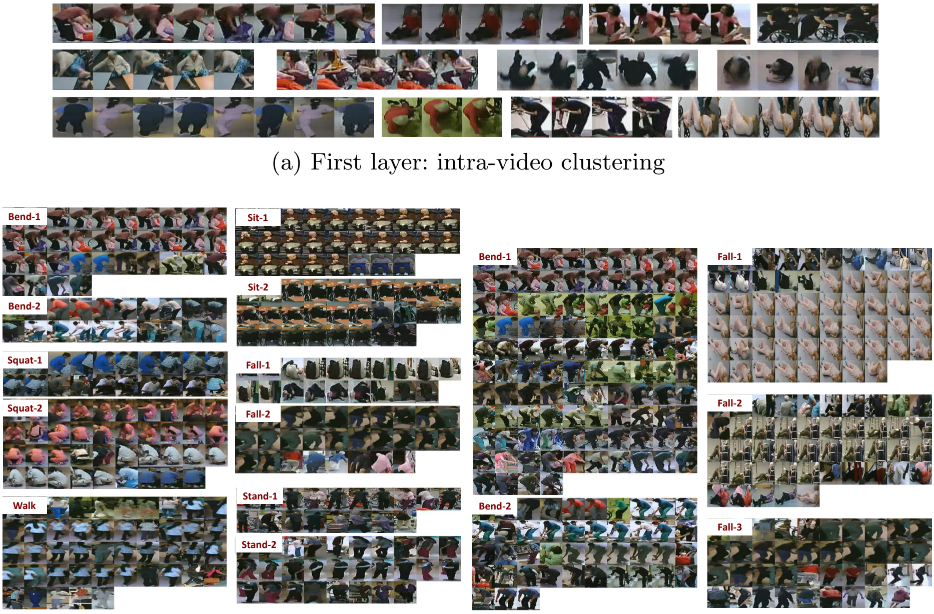
3.1 First Layer: Intra-Video Clustering

The first layer finds highly homogeneous action clusters for each video independently. Action examples in a video naturally form into multiple spatio-temporal clusters: examples within a small spatio-temporal volume tend to be consistent in appearance. Based on this intuition, we define the similarity between examples as an integration of appearance, spatial and temporal cues.

For the appearance similarity, we use a recently proposed exemplar-SVM based metric [28]. An exemplar SVM detector is trained for each positive example, and negative examples are randomly sampled from all video frames excluding the regions correspond to person. We use HOG as the feature descriptor. For each example, we run the detector on all other examples of the same action class in the video. We consider the top K scoring detections. The appearance similarity between a pair of examples i and j is defined as $d(i,j) = s(I_i, I_j)$, where I_i denotes the indices of the examples which are selected as the top K firings by the detector i. s measures how many times the detectors i and j are fired on the same window.

For the spatial similarity, we use the Euclidean distance between a pair of examples. The temporal similarity is defined as the number of frames between a pair of examples. We integrate these three similarities into the Medoid Shift clustering framework [32].

Fig. 3 (a) shows a visualization of several example clusters. Note that, due to our strategy of discriminative clustering and incorporating spatio-temporal relations between examples, most of the examples in the same cluster correspond to the same person in neighboring frames and are highly consistent in appearance.



(a) First layer: intra-video clustering

(b) Second layer: inter-video clustering (c) Third layer: action detector training

Fig. 3. Sample clusters in each layer. The first layer clusters actions within the same video. The second layer clusters actions in between videos. Finally, the third learns strong action detectors for further processing. Please refer to the text for details.

3.2 Second Layer: Inter-Video Clustering

We have obtained a large collection of atomic clusters, where each cluster contains highly consistent examples from the same video. The next step is to merge consistent atomic clusters from different videos. This step also relies on the same discriminative clustering scheme.

We train a linear SVM for each atomic cluster, where we use all examples in the atomic cluster as positive examples, and negative examples are randomly sampled from all video frames excluding the regions corresponding to a person. Similar to the intra-video clustering scheme, we run the detectors on all other examples of the same action class. Then we compute the affinity matrix, where the (i, j) entry of the matrix denotes the appearance similarity $d(i, j)$ (defined in the previous section) between atomic clusters i and j.

In this layer, we only use appearance cues to measure the similarity between pairs of atomic clusters. Once we have the affinity matrix, we do another level of clustering via the standard affinity propagation [33]. In this way, consistent atomic clusters from different videos are merged into one cluster. These clusters are used as our initial set of action primitives. Visualizations of example clusters in the second layer are shown in Fig. 3 (b).

3.3 Third Layer: Action Primitive Refining and Detector Training

The first two layers of our clustering framework automatically partition the person instances in each action class into a set of action primitives that are consistent in appearance, space and time. Now our goal is to train action detectors that can simultaneously recognize the action and localize the person who is performing the action. Modeling the action primitives that corresponds to the subcategories of the original action class will significantly reduce the intra-class variations and improve the detection performance. However, including noisy action primitives can cause the detector to become unreliable and thus hurt the action detection performance. In this work, we train the action detectors in the latent SVM framework, which iteratively learn the action detectors and refine the action primitives.

Object detector learning: We learn a multi-component action detector based on the DPM mixture model [6], where the mixture components are initialized by the action primitives discovered through our multi-layer clustering algorithm. We treat the action primitive labels as latent variables and allow them to refine during the latent step. Note that in standard DPM framework, the mixture components are initialized according to the examples' aspect ratios. However, the aspect ratio heuristic does not generalize well to a large number of subcategories, and thus often fails to provide a good initialization.

Action primitive pruning: There is no guarantee that all of the action primitive templates are discriminative. Weak templates can potentially put negative effects on detection results. We introduce a procedure to prune the templates that are not discriminative. We quantify this criterion with the average precision measure of action detection. We compute a precision-recall curve for each action primitive template; if its average precision is less than a threshold (0.5), we remove it. We compute precision with all positive examples, and a subset of 500 negative examples. The surviving action primitives are again used to initialize the multi-component action detector.

In our experiments, we used two iterations, as most good action primitives did not need more to converge (i.e. stop changing). We visualize the person examples in several sample action primitives in Fig. 3 (c).

4 Multi-Level Event Model

Our goal is to learn an event model that jointly considers persons' actions and action primitives, as well as the spatio-temporal interactions between them. We start with an example (Fig. 1) that illustrates modeling a fall event in a nursing home surveillance video. This scene includes a few actions like squat, fall and stand. Each action is fine-grained, represented by a certain action primitive, e.g. squat facing right, fall and sit on the floor, stand to push wheelchair, etc. We believe recognizing these actions helps to determine the event label, since it is common to find persons squatting, falling, standing, or running in a fall scene. Furthermore, the spatio-temporal interactions between these persons could also provide valuable information. For example, "losing balance", "falling on the floor", "squatting besides the fallen person" and "pushing wheelchair toward the fallen person" are in strict temporal ordering and form typical spatial patterns. We explicitly formulate spatio-temporal interactions in our model. Note that the interactions are between action primitives, rather than basic action categories. This is to remove ambiguity in generic actions – a fallen person sitting on the floor is likely interacting with a person standing to push a wheelchair, instead of a person standing still. Using action primitives enables us to discover these fine-grained cues for better video understanding.

To use our model in video recognition tasks, we employ the follow pipeline. During training, we have labeled frames where the event labels and action labels are provided, and we discover action primitives as described in Section 3. The frames are then represented by the multi-level event model, which is further learned in a max-margin framework to recognize events. During testing, we are given a test frame and we would like to decide the event label. We run our action primitive detectors to obtain candidate person detections, and reason about the event label from the detected actions, action primitives, spatio-temporal interactions, as well as the learned event model. Next, in Section 4.1, we formulate our multi-level event model. We then introduce the max-margin learning in Section 4.2.

4.1 Formulation

We first describe the multi-level labeling. Each video frame $\mathbf{x}$ is associated with an event label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of all possible event labels. Each person is associated with two labels: basic level action and action primitive. We use $\mathcal{H}$ and $\mathcal{Z}$ to denote the sets of all possible action and action primitive labels, respectively.

We encode three types of temporal information in our model: co-occurrence, before and after. We say that an action co-occurs with a video frame if the action takes place in the same temporal segment as that frame. Otherwise, the action is before or after the video frame. In our experiments, we consider the actions detected on that video frame as co-occurring. The before (or after) actions are those detected at most 20 sampled frames[1] before (or after) the given frame. We ignore the actions further away when modeling the current video frame. We denote a type of temporal information as $t \in \mathcal{T}$, where t equals to c, b and a representing co-occurrence, before and after, respectively.

To interpret $\mathbf{x}$ with the multi-level event representation, we find a candidate person x_i^t in each temporal segment t, and for each basic action category $i \in \mathcal{H}$, where $\mathcal{H}$ is the set of action labels. In our experiments, the candidate person x_i^t is simply set as the highest responding detection for action i in temporal segment t. However, we could easily extend our model to perform latent search over a set of candidate detections. Note that x_i^t is also associated with an action primitive label $z_i^t \in \mathcal{Z}$ with $\mathcal{Z}$ denotes all possible action primitives.

[1] In the nursing home dataset, 20 sampled frames account for roughly 10 seconds.

We now define the score of interpreting a video frame $\mathbf{x}$ with the multi-level event representation as:

$$F_\theta(\mathbf{x}, y) = \sum_{i \in \mathcal{H}, t \in \mathcal{T}} \alpha_{y,i,t} \cdot x_i^t + \sum_{i \in \mathcal{H}, j \in \mathcal{H}, t \in \mathcal{T}} \beta_{y,z_i^t,z_j^t,t}^\top \cdot [x_i^c, x_j^t, d_{ij}^t] \qquad (1)$$

where we use x_i^t interchangeably to denote the feature extracted from the bounding box of the person x_i^t. In our model, we set x_i^t as the scalar output of the action primitive detector for computational efficiency. Besides, $\theta = [\{\alpha\}, \{\beta\}]$ are the model parameters to be learned in the max-margin framework. We describe in detail each component in Eq. 1 in the following.

Unary event-action potential $\alpha_{y,i,t} \cdot x_i^t$: This potential captures the compatibility between the event y of the frame and each action i taking place in a given temporal segment. $\alpha_{y,i,t}$ is a scalar parameter that weights action i in temporal segment t for event y – high weights indicate discriminative actions.

Pairwise action primitive potential $\beta_{y,z_i^t,z_j^t,t}^\top \cdot [x_i^c, x_j^t, d_{ij}^t]$: This potential captures the compatibility between the event and pairs of action primitives. We fix the first person x_i^c to perform a co-occurring action since we target on modeling the current video frame. The second person x_i^t could be in any temporal segment to interact with person x_i^c. The term d_{ij}^t is a spatial feature computed based on the relative position and scale of person x_j^t's bounding box w.r.t. person x_i^c's bounding box. Note that d_{ij}^t is with respect to the second person's temporal segment t, which could be co-occurrence, before or after. Details of the spatial feature will be introduced in the following.

A straightforward way is to consider the interaction between every pair of action primitives. However, the model will become intractable and including irrelevant interactions will have negative effects on the event model. To handle this problem, we only consider a sparse set of interactions, by removing action primitive pairs that are infrequent (fewer than ten appearances) in the training data. For each selected action primitive pair, we follow [28] to extract the spatial feature. We start by fitting a two component Gaussian mixture model (GMM) to (the bounding boxes of) the pairs of action primitives. The GMM helps us to model various scale and spatial aspects of the action primitive pair. Moreover, we can produce a hypothesis for a bounding box by conditioning the learned GMM on the bounding box of a contextual person. We use the GMM output as the spatial feature. Formally, d_{ij}^t is the GMM score for person x_j^t's bounding box conditioned on person x_i^c's bounding box, where the GMM is trained for the action primitive pair (z_i^c, z_j^t).

Note that this pairwise potential accounts for spatio-temporal interactions between action primitives. The parameter $\beta_{y,z_i^t,z_j^t,t}$ identifies discriminative spatio-temporal interactions by assigning high weights.

4.2 Learning

We now describe how to learn the multi-level event model for video event recognition. Given a set of labeled training video frames $\{\mathbf{x}_i, y_i\}_{i=1}^N$, we would like to train the model parameters θ that tend to produce the correct event label for a new test video frame. A natural way of learning the model is to adopt the multi-class SVM formulation [34] as follows:

$$\min_{\theta,\xi\geq 0} \frac{\lambda}{2}||\theta||^2 + \sum_{i=1}^{N} \xi_i, \quad \text{s.t. } F_\theta(\mathbf{x}_i, y_i) - F_\theta(\mathbf{x}_i, y) \geq 1 - \xi_i, \quad \forall i, y \neq y_i \tag{2}$$

where $\xi = \{\xi_i\}_{i=1}^N$ are the slack variables to allow soft margin, and λ is a trade-off parameter. The constraint enforces that scoring a video frame $\mathbf{x}_i$ with the ground-truth label y_i is marginally larger than that with any other label $y \neq y_i$. The objective can be optimized using off-the-shelf solvers. In our experiments, we use a cutting-plane based solver implemented by [35].

5 Experiments

The focus of this work is on analyzing complex video events at multiple levels of granularity, including human actions and fine-grained primitives, spatio-temporal relations among multiple people and over-arching scene-level events. This type of structure widely exists in realistic multi-person scenes with rich social interactions. We demonstrate the effectiveness our approach with a challenging real-word application: fall detection in long-term care facilities. We have collected a large dataset of surveillance video footage from a nursing home facility – un-choreographed activity that contains substantial intra-class variation in action categories, and a natural setting to verify the efficacy of modeling complex activity structures.

5.1 Video Event Recognition

Understanding video events performed by multiple people has drawn lots of attention recently. However, the standard benchmark datasets on multi-person (group) activities (e.g. [36, 4]) are usually limited to pedestrian activities, such as walking together, talking, queueing, etc. In this work, we have collected a new challenging dataset for understanding multi-person activities in surveillance videos. Our dataset contains a diverse set of actions and primitives with large intra-class variations and thus presents lots of challenges in action recognition and localization. Focusing on the videos containing falls, this dataset naturally contains a rich set of realistic social interactions that form interesting spatio-temporal structures (e.g. squat beside a fallen person, lose balance, push wheelchair towards a person, etc). In the following, we first introduce the details of the dataset and experimental settings and then report the results.

Nursing Home Event Dataset: Our dataset consists of 125 video sequences (**in total 8 hours**) captured from fixed surveillance cameras mounted in a variety of rooms of a nursing home, including dining rooms, living areas, and corridors. Videos are recorded at 640 by 480 pixels at 24 frames per second. See Fig. 4 for example frames from the dataset.

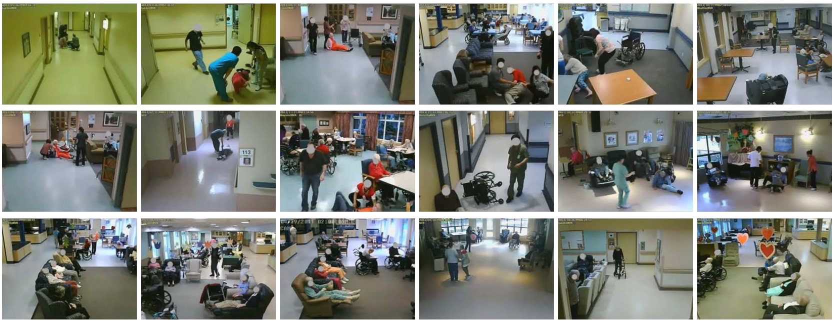


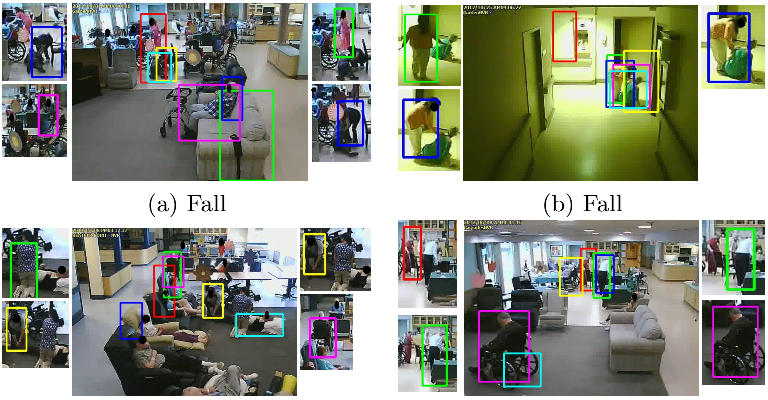
Fig. 4. Nursing Home Event Dataset. Our dataset contains 125 video sequences with six actions: walking, standing, sitting, bending, squatting, and falling. There are two event labels including fall (shown in the first two rows) and non-fall (shown in the last row). These video sequences are collected from a real-world nursing home surveillance project.

	Unary		Unary+Pairwise		
Model:	DPM	Primitive	Spatial	Temporal	Full model
AP (in%):	62.5	66.1	68.1	68.4	**68.6**

Table 1. Video event recognition performance on the nursing home dataset.

We split the dataset by using 104 short video clips for training (3 mins on average), and 21 relatively longer video clips for testing (8 mins on average). We annotated a subset (34769 frames) of all the frames in the training videos. Note that in this type of surveillance footage, it is common that there are no persons (or only static persons) appearing in the camera view over a relatively long period of time. Thus we skip annotating these frames. Our annotations include bounding boxes around the true location of the people in each frame (in the subset), action labels for each person, as well as the per-frame event labels. We define six action classes: walking, standing, sitting, bending, squatting, and falling, and two event classes: fall and non-fall. In order to remove redundancy, we sample 1 in every 10 frames for evaluation.

Baselines: We have designed four baselines to compare with our full model. The first DPM baseline runs DPM based action detectors [6], and detects actions for each video. The detection scores are then used in the unary model of Eq. 1, disregarding the temporal term t. Note that this method shares a similar principle to Action Bank [37]. The second baseline is the same as the first, but using the proposed action primitive detectors instead. We combine the action primitive detectors with the unary model of Eq. 1, which results in a hierarchical structured (event and actions) model. This is an example of a standard structured model for action recognition. The third spatial baseline uses the spatial information only by disregarding the other temporal segments (i.e. setting $\mathcal{T} = \{c\}$). Finally, the last temporal baseline considers only the temporal information by removing the spatial feature d_{ij}^t from our full model. Note that the



(a) Fall

(b) Fall

(c) Fall

(d) Non-fall

Fig. 5. Visualizations of our result. We select four frames, and show the detected actions in each frame. Each bounding box is marked by a color, which denotes the predicted action class. We use green, red, purple, blue, yellow and cyan to represent walking, standing, sitting, bending, squatting, and falling respectively. The actions used in the before and after segments are listed in the left and right of the frame, respectively. Our model captures the spatio-temporal interactions among these actions to predict the event labels (as captioned below each frame): the first three frames are all believed to be fall scenes with the last being non-fall.

spatial baseline, the temporal baseline, and our full method learn with both the unary event-action potential and the pairwise action primitive potential.

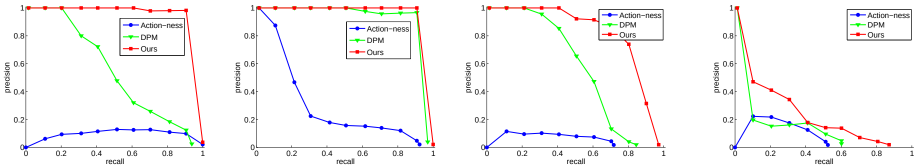
For a fair comparison, we use the same solver for learning all these methods. The trade-off parameter λ in Eq. 2 is simply set as 1 for all experiments. The recognition performance is measured by average precision (AP) of fall detection.

Results: The results are listed in Table 1. We first compare the two baselines using unary models only. The action primitive based baseline outperforms the DPM based method. This validates the usage of our learned action primitives. Furthermore, adding the pairwise model to the unary model improves the overall recognition performance. Specifically, our full model outperforms all other baseline methods including the temporal model and the spatial model. This result verifies that the pairwise potentials capture useful spatio-temporal information for recognizing video events. We have proposed a unified framework that builds over low-level action primitives and mid-level actions to analyze high-level video events. Intuitively, one can model spatio-temporal interactions among action primitives to capture useful cues. The result shows that the unified framework can be effective on a challenging dataset, and performs better than standard approaches using Action Bank-type representations and other structured models.

Visualizations of our results are shown in Fig. 5, which shows that our model captures spatio-temporal interactions between action primitives to reason about the event label. For example, in Fig. 5 (a), the nurse in red stood to push the wheelchair when she saw the fallen person, and then walked out to call for help.

Model:	Action-ness	DPM	Ours
mAP (in%):	37.1	49.8	**55.8**

Table 2. Action localization performance on the UCF-Sports dataset. Action-ness stands for the first baseline that runs on all the action bounding boxes, and DPM is the second baseline that trains a DPM for each action class.



(a) horse riding (b) lifting (c) swing: high bar (d) swing: pommel horse

Fig. 6. Precision-recall curves of four sample action classes on the UCF-Sports dataset. These action classes have obvious action sub-categories and thus benefit from our action primitive based model.

Another man in a black shirt stayed right beside to help the fallen person by performing a series of actions (bending-squatting-bending). These are obvious cues for a fall scene. Moreover, in the scene of Fig. 5 (b), a nurse walked toward the fallen person, and then bent to help. Note that the actions detected in the before and after segments compensate for the noisy detections in the video frame, and together are used to interpret this as a “fall” scene. In Fig. 5 (d), we correctly recognize this non-fall scene although there is a false detection of a falling action. This is because this scene has no spatio-temporal interactions between bending, squatting and falling that are commonly seen in fall scenes.

5.2 Action Localization

Higher-level modeling of structured human activities is aided by accurate action localization. In order to verify the performance of our action primitives, we use them for action localization on the popular UCF-Sports dataset [38].

Dataset: The UCF-Sports dataset consists of 150 broadcast videos from 10 action classes ranging from diving, golf swinging, kicking, lifting, horse riding, running, skating, swinging (on the pommel horse and on the floor), and swinging (at the high bar), to walking. We follow the training/testing split proposed in [23] and use 103 videos for training and 47 for test. We use the ground-truth bounding annotations provided in the training data.

Baseline: We compare our action primitive based action localization model with the following baselines. The first baseline is an “action-ness” detector (c.f. [39]) that is simply a DPM trained on all bounding boxes without considering the action class. The second baseline executes a standard DPM detector for each action class. As for our model, we follow the steps described in Section 3 to generate action primitives. After learning, an average of 4 action primitives are discovered for each action class.

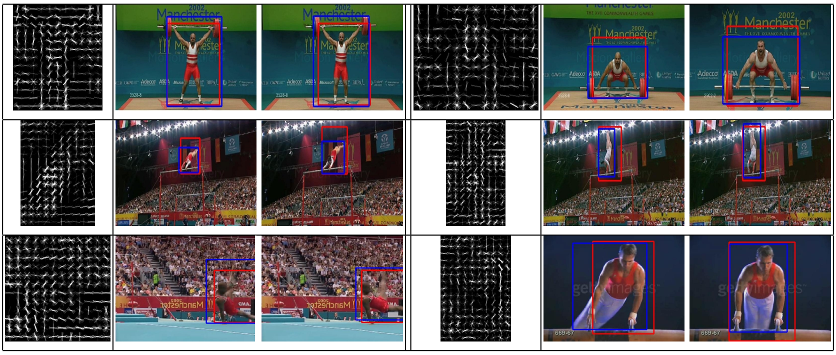


Fig. 7. Sample visualization results of our localization model. Each row shows two sample action primitives of an action class, e.g. lifting, swing: high bar, and swing: pommel horse (from top to down). For each action primitive, we visualize the learned model on two sample video frames, where the highest responding detections are shown in red rectangles. For comparison, we also plot the ground-truth bounding boxes in blue rectangles.

For performance evaluation, we run each compared model on the test videos. We assume that the action recognition is perfectly done so we use the corresponding action detector for each action class, for the DPM model and ours. We collect the detector responses on each frame, and measure the mean average precision according to the PASCAL VOC criterion [40].

Results: The mAP results are reported in Table 2, which shows that our model outperforms all the baselines. We have also selected sample action classes and plotted the precision-recall curve in Fig. 6. These results again validate the utility of action primitives in localizing actions. We visualize sample localization results in Fig. 7. As can be seen, the action primitives are well-localized in many instances. Detailed, accurate localization of this form can permit the type of high-level activity reasoning that our full model can produce.

6 Conclusion

We presented an algorithm for learning a multi-level representation for the actions of people in a scene. In order to address the intra-class variation of an action category, we developed a data-driven approach to discover action primitives. These action primitives model specific appearance, viewpoint, and temporal stage variants of an action category. An algorithm for automatically discovering these primitives from only action-level supervision was presented, based on clustering and discriminative selection of primitives. A multi-level model for the actions of people in a scene was built around these primitives, allowing us to model detailed inter-relations among action primitives. Empirical results showed that these primitives permit effective localization of actions, improved recognition of human actions, and a detailed explanation of human behaviour in an entire scene-level event.

References

1. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: NIPS. (2010)
2. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: CVPR. (2012)
3. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: ECCV. (2012)
4. Amer, M.R., Xie, D., Zhao, M., Todorovic, S., Zhu, S.C.: Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In: ECCV. (2012)
5. Ramanathan, V., Yao, B., Fei-Fei, L.: Social role discovery in human events. In: CVPR. (2013)
6. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. T-PAMI **32** (2010) 1672–1645
7. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. T-CSVT (2008)
8. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR. (2004)
9. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: CVPR. (2010)
10. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR. (2011)
11. Jain, A., Gupta, A., Rodriguez, M., Davis, L.S.: Representing videos using mid-level discriminative patches. In: CVPR. (2013)
12. Wang, H., Kläser, A., C.Schmid, Liu, C.L.: Action recognition by dense trajectories. In: CVPR. (2011)
13. Raptis, M., Kokkinos, I., Soatto, S.: Discovering discriminative action parts from mid-level video representations. In: CVPR. (2012)
14. Tian, Y., Sukthankar, R., Shah, M.: Spatiotemporal deformable part models for action detection. In: CVPR. (2013)
15. Shugao Ma, Jianming Zhang, N.I.C., Sclaroff, S.: Action recognition and localization by hierarchical space-time segments. In: ICCV. (2013)
16. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden markov model. In: CVPR. (1992)
17. Moore, D., Essa, I.: Recognizing multitasked activities from video using stochastic context-free grammar. In: AAAI. (2002)
18. Bobick, A., Wilson, A.: A state-based technique for the summarization and recognition of gesture. In: ICCV. (1995)
19. Bregler, C.: Learning and recognizing human dynamics in video sequences. In: CVPR. (1997)
20. Médioni, G., Cohen, I., Brémond, F., Hongeng, S., Nevatia, R.: Event detection and analysis from video streams. T-PAMI **23** (2001) 873–889
21. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: ICCV. (2007)
22. Kläser, A., Marszałek, M., Schmid, C., Zisserman, A.: Human focused action localization in video. In: International Workshop on Sign, Gesture, Activity. (2010)
23. Lan, T., Wang, Y., Mori, G.: Discriminative figure-centric models for joint action localization and recognition. In: ICCV. (2011)
24. Tran, D., Yuan, J.: Max-margin structured output regression for spatio-temporal action localization. In: NIPS. (2012)

25. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. (2010)
26. Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Discovering primitive action categories by leveraging relevant visual context. In: ECCV Workshop on Visual Surveillance. (2008)
27. Hoai, M., Zisserman, A.: Discriminative sub-categorization. In: CVPR. (2013)
28. Lan, T., Sigal, L., Raptis, M., Mori, G.: From subcategories to visual composites: A multi-level framework for object detection. In: ICCV. (2013)
29. Gu, C., Ren, X.: Discriminative mixture-of-templates for viewpoint classification. In: ECCV. (2010)
30. Todorovic, S., Ahuja, N.: Learning subcategory relevances for category recognition. In: CVPR. (2008)
31. Gu, C., Arbelaez, P., Lin, Y., Yu, K., Malik, J.: Multi-component models for object detection. In: ECCV. (2012)
32. Sheikh, Y.A., Khan, E.A., Kanade, T.: Mode-seeking via medoidshifts. In: ICCV. (2007)
33. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science (2007)
34. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. JMLR **2** (2001) 265–292
35. Do, T.M.T., Artieres, T.: Large margin training for hidden markov models with partially observed states. In: ICML. (2009)
36. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatial-temporal relationship among people. In: International Workshop on Visual Surveillance. (2009)
37. Sadanand, S., Corso, J.J.: Action Bank: A high-level representation of activity in video. In: CVPR. (2012)
38. Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH: A spatial-temporal maximum average correlation height filter for action recognition. In: CVPR. (2008)
39. Alexe, B., Deselares, T., Ferrari, V.: What is an object? In: CVPR. (2010)
40. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. IJCV **88** (2010) 303–338