Integrating Multi-Stage Depth-Induced Contextual Information For Human Action Recognition and Localization

Bingbing Ni, Yong Pei
Advanced Digital Sciences Center
Singapore 138632
bingbing.ni@adsc.com.sg
pei.yong@adsc.com.sg

Zhujin Liang, Liang Lin
Sun Yat-Sen University
China 510006
alfredtofu@gmail.com
linliang@ieee.org

Pierre Moulin
UIUC
IL 61820-5711 USA
moulin@ifp.uiuc.edu

***Abstract*— Human action recognition and localization is a challenging vision task with promising applications. To tackle this problem, recently developed commodity depth sensor (*e.g.*, Microsoft Kinect) has opened up new opportunities with several developed human motion features based on depth image for action representation. However, how depth information can be effectively adopted in the middle or high level representation in action detection, in particular, the depth induced three dimensional contextual information for modeling interactions between human-human, human-object and human-surroundings has yet been explored. In this paper, we propose a novel action recognition and localization framework which effectively fuses depth-induced contextual information from different levels of the processing pipeline for understanding various interactions. First, depth image is combined with grayscale image for more robust human subject and object detection. Second, three dimensional spatial and temporal relationship among human subjects or objects is represented based on the combination of grayscale and depth images. Third, depth information is further utilized to represent different types of indoor scenes. Finally, we fuse these multiple stage depth-induced contextual information to yield an unified action detection framework. Extensive experiments on a challenging grayscale + depth human action detection benchmark database demonstrate the effectiveness of the depth-induced contextual information and the high detection accuracy of the proposed framework.**

I. INTRODUCTION

Video based human activity analysis has attracted significant interest in recent years given its promising applications such as smart video surveillance, video event retrieval and video based human computer interaction (HCI). In a realistic setting, it is required to not only recognize but also localize certain action of interest in the video sequence spatially and temporally. Action recognition and localization is a difficult problem due to the individual variations of people in posture, motion and clothing, view angle changes and camera motions, illumination changes, occlusions and self-occlusions, and the complex and cluttered background.

Various methods have been proposed during decades for action recognition. These methods include motion templates [1], silhouettes [7], space-time shapes [6], trajectories [13], etc. Spatio-temporal interest points (STIPs) [10], [8], [4] are nowadays most widely used to effectively represent motion information. Typically, interesting parts are extracted from cuboids surrounding local maxima of spatio-temporal corners, *i.e.*, locations of videos which exhibit strong variations of intensity both in spatial and temporal directions. Then, bag-of-words method is used for pooling the local features to form a global representation. More recently, the method based on dense interesting point trajectories [14] extracted from video sequences has shown its significant improvement over previous art in terms of action recognition accuracy.

Recent emergence of depth sensor (*e.g.*, Microsoft Kinect) has made it feasible and economically sound to capture in real-time not only the color images, but also depth maps with appropriate resolution (*e.g.*, 640×480 in pixel) and accuracy. It can provide three-dimensional structure information of the scene as well as the three-dimensional motion information of the subjects/objects in the scene. Therefore the motion ambiguity of the color camera, *i.e.*, projection of the three dimensional motion onto the two-dimensional image plane, could be bypassed. Based on the depth image, many methods have been proposed for improving the human action detection performance. These methods include using the $3D$ point cloud [11] and the $3D$ joint position data for action representation and classification. In these methods, depth information is mainly utilized for representing human subjects' motion information, *e.g.,*, the movement of the $3D$ joints [15] or three dimensional human pose and shape [11]. In this work, we show that when considering recognizing and localizing action which involves interactions among human-human, human-object and human-surroundings, depth induced spatial-temporal contextual information can play an even more important role. Integrating various contextual information for action recognition has recently received much attention [9]. However, as previous art only utilizes conventional video cameras, these interaction relationship and contextual information are measured in $2D$ image domain. Using conventional camera, projection of $3D$ position to $2D$ images loses the absolute spatial coordinates and thus ambiguity arises. For example, objects with large displacement in the depth direction could be very close in the $2D$ image. However, using depth image resolves this problem. Therefore accurate modeling of three dimensional spatial temporal relationship is very critical for accurate action detection. We show that using the additional depth image, richer three dimensional spatial temporal information can be adopted for more accurate interaction and context

modeling.

In this paper, we propose a novel action recognition and localization framework which effectively fuses depth-induced contextual information from different levels of the processing pipeline for boosting the action (in particular, interactions between human-human, human-object, or human-surroundings) detection performance. First, depth image is combined with grayscale image for more robust human detection. Second, three dimensional spatial and temporal relationship among human-human, human-object, or human-surroundings are modeled based on the combination of greyscale and depth image. Third, depth information is further utilized to represent different types of indoor scenes. Finally, we fuse these multiple stage depth information to yield an unified action detection framework. Extensive experiments on a recent challenging grayscale + depth action detection database demonstrate the effectiveness of depth fusion in multiple processing stages and significant action detection accuracy improvement over the previous art.

The rest of this paper is organized as follows. First, we discuss some related works in Section II. Section III presents the proposed multi-stage depth information fusion framework for action recognition and localization. Extensive experimental results on a challenging benchmark dataset on grayscale + depth action detection are given in Section IV. Section V concludes the paper with discussions on future works.

II. RELATED WORKS

Li et al. [11] presented a method to recognize human actions from sequences of depth maps. An action graph is employed to model explicitly the dynamics of the actions and a bag of $3D$ points to characterize a set of salient postures that correspond to the nodes in the action graph. This technique has been successfully applied in recognizing a set of actions such as waving hands, jumping etc, for the purpose of human computer interaction.

Ni et al. [12] developed two color-depth fusion schemes for feature representation from the most representative feature representation methods in human action recognition. They first extend the spatio-temporal interest points methods (STIPs) into a depth-layered multichannel representation; then they augment the motion history images (MHIs) with two depth change induced motion history channels. Superior performances are gained by fusing color and depth information for human daily activity recognition.

It is generally agreed that knowing the $3D$ joint position is helpful for action recognition. Based on the depth data and the estimated $3D$ joint positions, Yuan et al. [15] proposed a new translational invariant local occupancy pattern feature, associated with each $3D$ joint as the depth appearance of this $3D$ joint. They defined a particular conjunction of the features for a subset of the joints, indicating a structure of the features as *actionlet*. A data mining solution to discover discriminative actionlets is also proposed. Then an action is represented as a linear combination of actionlet ensemble. Extensive experimental results show that the proposed method is able to achieve significantly better recognition accuracy than the state-of-the-art methods.

Note that our work is not a counterpart to these previous works. Rather, our proposed framework can be regarded as complementary to these works. While the previous methods only consider some low level representation *e.g.*, motion information or $3D$ joint position using the depth image, our work focuses on effective modeling of the depth-induced contextual information, which is high level representation. As motion and contextual information are both important to action analysis, we believe combining our proposed framework with previously developed depth based motion features can further boost the action detection performance.

III. METHODOLOGY

A. Overview of Our Method

Our basic idea in this work is to integrate multiple stage depth-induced contextual information for detecting human activities that involves interactions between human-human, human-object or human-surroundings. Figure 1 gives an illustration of our proposed processing pipeline for action recognition and localization. Depth information at various processing stages is utilized and integrated in the following way. First, we detect human key pose and object of interest in every input frame of the grayscale image sequence and the corresponding depth images are used to filter out false detections. These human key pose and object detections are afterwards spatially and temporally matched throughout frames into *tracklets* by reasoning the motion constraints in both grayscale and depth channel. As a by-product, invalid detections without sufficient temporal durations are further filtered out at this step. In the next step, we model the three dimensional spatial-temporal interaction/contextual attributes using combined grayscale and depth information. In the meantime, depth information is utilized for classifying the indoor scenes into different scene categories. In the final step, the obtained spatial-temporal interaction attributes, key pose of the tracklets and the scene classification results are fused together via a Bayesian network for action recognition and localization.

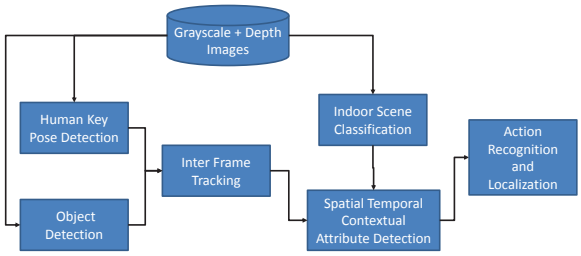

Fig. 1. Overview of the proposed action detection framework.

B. Depth-Aided Human/Object Detection and Tracking

Although there exist various complex representations of human actions, human can easily recognize what a person is doing even by looking at a single frame without examining

the whole sequence. Therefore, in this study, we use human key pose as a primitive representation for human actions. The advantages of this representation are as follows: 1) key pose based representation is more compact and therefore it reduces the computational complexity; and 2) it is robust to the variance in execution styles of the same action.

Human key poses are represented by histogram of oriented gradient (HOG) features [3] extracted from grayscale image. In this work, we use 8×8 blocks for calculating the gradient histogram, as in the implementation [5]. Neighborhood blocks are considered when pooling the gradients for the current block. Different actions are associated with different key poses. To obtain representative key poses, we cluster the given ground truth annotations of human subjects into different groups via K-means algorithm and each cluster is used to represent a key pose prototype. Note that the same key pose can be shared among multiple actions. For key pose detection, we train linear SVM detector based on HOG features [3] using the training samples associated with each cluster (group, or key pose prototype). In this work, we cluster the training samples into 28 types of key poses. Similarly, we also detect objects of interest using the above method. Each detection instance $\mathbf{x}$ is assigned to a key pose type $k, k = 1, 2, \cdots, K$, *i.e.*, K key poses, via nearest cluster center assignment.

Using the grayscale image alone for human key pose or object detection is not prune to false alarms. Depth based constraints can be used to effectively remove false detections. In this work, we utilize two heuristic depth based constraints. We define these constraints in the following. First, the area to median depth ratio for a human subject should be within a certain range. Suppose for human subject detection $\mathbf{x}$, its area and median depth are denoted as $Area(\mathbf{x})$ and $d_m(\mathbf{x})$, respectively. Then this constraint can be formally denoted as:

$$r_l \leq r(\mathbf{x}) = \frac{Area(\mathbf{x})}{d_m(\mathbf{x})} \leq r_u, \tag{1}$$

where $r(\mathbf{x})$ denotes the area to median depth ratio and r_l and r_u denotes the lower and upper bounds, respectively. Second, the median human body depth value should be smaller (*i.e.*, nearer to the camera center) than the depth values surrounding the human body in the horizontal direction. To enforce this constraint, we denote $d_m(\mathbf{x}-\mathbf{g})$ and $d_m(\mathbf{x}+\mathbf{g})$ as the median depth value of the image stripe located on the left and right side of the human detection bounding box, respectively, and the constraint can be expressed as:

$$d_m(\mathbf{x}-\mathbf{g}) > d_m(\mathbf{x}), d_m(\mathbf{x}) < d_m(\mathbf{x}+\mathbf{g}). \tag{2}$$

The diagram illustration of these two constraints is given in Figure 2. All the parameters can be estimated from the training data. In the experiment, we note that large number of false detections could be filtered out using these two depth based constraints.

Once the candidate per frame subject/object detections are obtained, we temporally tracked them into human or object sequences, named as *tracklet*. The tracklet extraction process

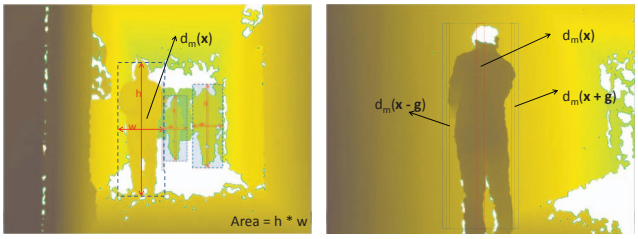

Fig. 2. Depth induced constraints for human detection. The left diagram is for illustrating (1) and the right is for (2).

is based on the pairwise detection matching over consecutive frames. First, we establish all the key pose or object detection matches between frame i and $i+1$. Matches that extend over several frames then form a motion trajectory of the key poses or objects. To match two detections $\mathbf{x}$ and $\mathbf{y}$ in consecutive frames, we calculate the matching score which is the weighted combination of the three dimensional spatial displacement between consecutive frames and the differences of detection bounding boxes as,

$$dist(\mathbf{x},\mathbf{y}) = \alpha\|\mathbf{p}(\mathbf{x}) - \mathbf{p}(\mathbf{y})\|_2 + \beta\|a(\mathbf{x}) - a(\mathbf{y})\|_2, \tag{3}$$

where $\mathbf{p}(\mathbf{x})$ denotes the three dimensional coordinate of the center of of the detection x, *i.e.*, $\mathbf{p}(\mathbf{x}) = (x, y, z)^T$. We impose that for any detection in frame i, there can be maximally one candidate match in frame $i+1$. Also, two detections are considered as un-matched if their $dist(\mathbf{x},\mathbf{y})$ is larger than some threshold value t_{dist}. Parameters α, β and t_{dist} can be set empirically based on the training data. To further remove possible noisy tracklets, we restrict the length L of any valid tracklet to be L_{min} and L_{max}. In this work, we set $L_{min} = 5$ and $L_{max} = 200$, which correspond to 0.2 to 8 seconds in duration. Key pose is a good indicator for action category, therefore, for each type of key pose k, we calculate its probability of representing action class j as $m(k, j)$. This value could be estimated empirically from the training dataset as:

$$m(k, j) = \frac{|\{\mathbf{x} : \mathbf{x} \in Pose(k) \bigcap \mathbf{x} \in Action(j)\}|}{|\{\mathbf{x} : \mathbf{x} \in Pose(k)\}|}. \tag{4}$$

Here $\mathbf{x} \in Pose(k)$ means the detection $\mathbf{x}$ belongs to key pose type k; $\mathbf{x} \in Action(j)$ means the detection $\mathbf{x}$ is contained in some instance of action class j. $|\{\}|$ denotes the number of elements in the set. Estimation is performed over all the training instances of key pose detections and the ground truth action bounding boxes. Therefore, each key pose prototype k can be represented as a C-dimensional vector (*i.e.*, the number of action classes is C) as $\mathbf{m}(k) = [m(k,1), m(k,2), \cdots, m(k,C)]^T, k = 1, 2, \cdots, K$, which indicates its action class association probability. For a tracklet T, we then calculate its class probability $\mathbf{f}_u$ by averaging the detections over the whole tracklet as:

$$\mathbf{f}_u(T) = \frac{1}{|\{\mathbf{x} \in T\}|} \sum_{\mathbf{x}} \mathbf{m}(k(\mathbf{x})). \tag{5}$$

Here $|\{\mathbf{x} : \mathbf{x} \in T\}|$ denotes the number of element in the tracklet T and $k(\mathbf{x})$ denotes the type of key pose for detection

$\mathbf{x}$. We refer $\mathbf{f}_u(T)$ as the unary attribute for tracklet T in our integrated formulation for action detection.

C. 3D Spatial-Temporal Contextual Attributes

It is widely recognized that modeling spatial-temporal contextual information is important for recognizing various interactions including human-human, human-object or human-surroundings [9]. Previous methods mostly utilize conventional $2D$ images therefore the contextual information can only be measured in $2D$. As depth image is available, spatial temporal contextual information for interaction can be measured directly in $3D$. In this work, we developed a set of three dimensional spatial temporal contextual attributes between tracklets of human key poses or objects, by explicitly utilizing the combined information from grayscale and depth images. The advantage of modeling $3D$ spatial temporal contextual interaction attributes between two tracklets is obvious. For example, the action *discussion of two people* requires that two human subjects standing nearby within a relatively fixed distance over a sequence of frames, and their relative speed should be approximately zero. However, the actual distance between two human subjects inferred from a conventional image could be ambiguous as their displacement in the depth direction is not considered. Using the additional depth image solves this problem trivially. More specifically, we define the following three dimensional spatial temporal interaction attributes between two tracklets T_1 and T_2:

1) **Relative** $3D$ **distance**: The relative distance between two tracklets is defined as the mean value of the three dimensional displacement between two temporally overlapping tracklets over their common frames as:

$$\mathbf{f}_d(T_1,T_2) = \frac{1}{|t_e - t_s + 1|}\sum_{i=t_s}^{t_e}(\mathbf{p}(\mathbf{x}_i^{T_1}) - \mathbf{p}(\mathbf{x}_i^{T_2})). \quad (6)$$

Here t_s and t_e means the start and end frame number of the temporally overlapping portion of tracklets T_1 and T_2. $\mathbf{x}_i^{T_1}$ denotes the human subject/object detection in i-th frame (with respect to the t_s) of the tracklet T_1. Relative distance is useful for identifying some types of interactions such as *discussion of two people*, when the $3D$ displacement between two human subjects has high discriminative capability. We discretize $\mathbf{f}_d$ into several ranges in each axis so that it can take a finite number of values.

2) **Relative** $3D$ **velocity**: The mean relative velocity between two overlapping tracklets over their common frames as:

$$\mathbf{f}_v(T_1,T_2) = \frac{1}{|t_e - t_s + 1|}\sum_{i=t_s}^{t_e}(\mathbf{v}(\mathbf{x}_i^{T_1}) - \mathbf{v}(\mathbf{x}_i^{T_2})). \quad (7)$$

Here $\mathbf{v}(\mathbf{x}_i)$ denotes the $3D$ velocity of the detection $\mathbf{x}$ in consecutive frames i and $i+1$, *i.e.*, $\mathbf{v}(\mathbf{x}_i) = \mathbf{p}(\mathbf{x}_{i+1}) - \mathbf{p}(\mathbf{x}_i)$. Relative $3D$ speed is useful for some types of interactions such as *A person unlocks an office and then enters it*, which typically involves a phase of fixed relative speed between human and door (approach the door), a phase of fixed zero relative speed (*i.e.*, unlock the door) and another phase of fixed relative speed (*i.e.*, enter). Similarly, we discretize $\mathbf{f}_v$ into finite number of values.

3) **Relative temporal ordering**: Temporal ordering relationship between two tracklets are also very important for distinguishing actions which involves certain temporal ordering pattern. For example, the action *A person tries to enter an office unsuccessfully* contains three sequential events: approach the door, try to unlock the door, and leave the door. We define three types of temporal ordering relationship between two tracklets, namely, *overlap*, *precede* or *succeed*. T_1 and T_2 is considered as *overlap* when the portion of their overlapping frames is significant as (we assume there are overlapping frames, otherwise the value is zero):

$$\frac{min(t_e(T_1),t_e(T_2)) - max(t_s(T_1),t_s(T_2)) + 1}{max(t_e(T_1),t_e(T_2)) - min(t_s(T_1),t_s(T_2)) + 1} > t_{ol}, \quad (8)$$

where t_{ol} is some empirically set threshold. T_1 is considered as preceding T_2 when $t_e(T_1) - t_s(T_2) < t_{pr}$ and vice versa. t_{pr} is some empirically set threshold.

We refer to $\mathbf{f}_o(T_1,T_2)$ as the temporal ordering attribute between tracklets T_2 and T_2, which is a three dimensional binary vector.

D. Depth Based Scene Classification

Knowing the type of the scene can also benefit action recognition. For example, the action *A person types on a keyboard* usually does not occur in an outside-office scene (*e.g.*, corridor). Similarly, the action *A person unlocks an office and then enters it* always occurs in the scene which contains a door. We note that when depth image is available, we can use the $3D$ geometric attributes for modeling the scene type. In particular, we can use $3D$ plane orientations. To model the scene geometric structure, we first transform depth image into normal map. Normal map is composed of values of each cell plane's (patch) $3D$ normal, which is a normalized vector that indicates the orientation of a plane. $3D$ normals can be directly computed by fitting a plane equation from the $3D$ points sampled from an image local patch which are assumed to be on the same plane. For each pixel in depth map, we use its 7×7 neighborhood pixels' $3D$ coordinates to fit a panel, and compute the plane's normal. After calculating the normals, we project the $3D$ plane directions onto the $2D$ image plane, and we further represent the scene by the histogram of four major orientations (up, down, left, right) and using the orientation histogram to classify the scene type by support vector machine [2]. In this work, we define 5 types of indoor scenes. Exemplar plane orientation map is given in Figure 3. We refer to $\mathbf{f}_s$ as the scene type attribute, which takes 5 possible values.

E. Action Detection

We use Bayesian network for action inference. Our probabilistic model integrates the unary attribute of tracklet, the

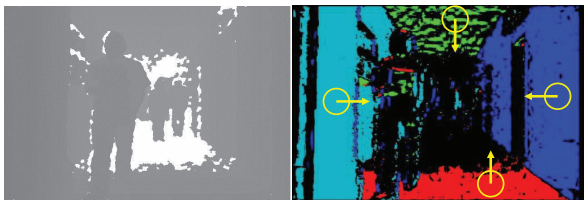

Fig. 3. Example scene type (depth image) and its corresponding projected normal map. Note that red, green, blue, cyan represent up, down, left, right, respectively.

spatial temporal interaction contextual attributes between tracklets, and the scene type attribute for action detection, as shown in Figure 4. Our joint probability model for an action a given two tracklet T_1 and T_2 and the scene type attribute $\mathbf{f}_s$ is given by:

$$\begin{aligned} & p(a|T_1,T_2,\mathbf{f}_s) \\ & = p(a|\mathbf{f}_u(T_1),\mathbf{f}_u(T_2),\mathbf{f}_d(T_1,T_2),\mathbf{f}_v(T_1,T_2),\mathbf{f}_o(T_1,T_2),\mathbf{f}_s) \\ & \sim P(a|\mathbf{f}_u(T_1))P(a|\mathbf{f}_u(T_2)) \\ & \times P(a|\mathbf{f}_d(T_1,T_2))P(a|\mathbf{f}_v(T_1,T_2))P(a|\mathbf{f}_o(T_1,T_2)) \\ & \times P(a|\mathbf{f}_s) \end{aligned} \quad (9)$$

Note that all random variables are discrete, and all attribute distributions are multinomial. We train the model using Expectation Maximization (EM) algorithm, and classify a novel video sequence by finding the action a_{opt} that maximizes $p(a|T_1,T_2,\mathbf{f}_s)$. Uniform priors are assumed. The action is localized by the smallest video volume that contains the tracklet T_1 and T_2.

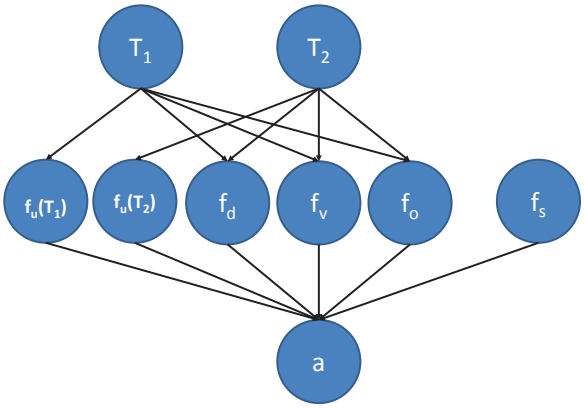

Fig. 4. Bayesian network for action inference.

IV. EXPERIMENT

We use the HARL 2012 competition dataset [16] for experiment. The goal of the HARL 2012 competition focuses on the complex human behavior involving several people in the video at the same time, on actions involving several interacting people and on human-object interactions. The goal is not only to classify activities, but also to detect and to localize them. The dataset is shot with two different cameras: a moving camera mounted on a mobile robot delivering grayscale videos in VGA resolution and depth images from a consumer depth camera (Kinect). The resolutions of both grayscale and depth image are 640×480 pixels. The dataset contains grayscale/depth videos (D1) showing people performing various activities taken from daily life (discussing, telphone calls, giving an item etc.) The dataset is fully annotated, where the annotation not only contains information on the action class but also its spatial and temporal positions in the video (bounding boxes). The dataset consists of 10 classes. Each of classes can be a normal activity, a human-human interaction or a human-object interaction, or a combination of the latter two types. These actions include: *DI: Discussion of two or several people*, *GI: A person gives an item to a second person*, *BO: An item is picked up or put down*, *EN: A person enters or leaves an office*, *ET: A person tries to enter an office unsuccessfully*, *LO: A person unlocks an office and then enters it*, *UB: A person leaves baggage unattended*, *HS: Handshaking of two people*, and *KB: A person types on a keyboard*, *TE: A person talks on a telephone*. In total, the training set has 305 action samples and the testing set has 156 samples.

We use the action detection performance evaluation metric defined in [16]. The goal is to measure a match between the annotated ground-truth and a result, *i.e.*, between: 1) a list G of ground truth actions $G^{v,a}$, where $G^{v,a}$ corresponds to the a-th action in the v-th video and where each action consists of a set of bounding boxes $G_b^{v,a}$ marked with the same class; and 2) a list D of detected actions $D^{v,a}$, where $D^{v,a}$ corresponds to the a-th action in the v-th video and where each action consists of a set of bounding boxes $D_b^{v,a}$ marked with the same class. The measure first finds the best match for an action in a list of potential candidate matches. It maximizes the normalized overlap between the two actions over all frames:

$$Recall(G,D) = \frac{\sum_v \sum_a IM(G^{v,a}, BM(G^{v,a}, D^v))}{\sum_v |G^v|}, \quad (10)$$

$$Precision(G,D) = \frac{\sum_v \sum_a IM(BM(D^{v,a}, G^v), D^{v,a})}{\sum_v |D^v|}. \quad (11)$$

Here the *Best Match* (BM) is defined by:

$$BM(X^{v,a}, Y^v) = argmax_{a'=1\cdots|Y^v|} \frac{2 \times Area(X^{u,a} \bigcap Y^{u,a'})}{Area(X^{u,a}) + Area(Y^{u,a'})}. \quad (12)$$

To give a formal expression for the *Is Matched* (IM) criteria, we abbreviate the ground truth action by $g = G^{v,a}$ and the detected action by $d = D^{v,a'}$. Furthermore, we denote by $g|d$ the set of bounding boxes of the ground truth action g restricted to uniquely the frames which are also part of detected action d. In a similar way, we denote by $d|g$ the set of bounding boxes of the detection action d restricted to uniquely the frames which are also part of ground truth action g. Then, $IM(g,d) = 1$ when the following conditions

are ALL satisfied:

$$\frac{Area(g\bigcap d)}{Area(g|d)} > t_{sr}, \frac{Area(g\bigcap d)}{Area(g|d)} > t_{ps},$$
$$\frac{NoFrames(g\bigcap d)}{NoFrames(g)} > t_{rt}, \frac{NoFrames(g\bigcap d)}{NoFrames(d)} > t_{pt},$$
$$class(g) == class(d) \quad (13)$$

where $Area(X)$ is the sum of the areas of the bounding boxes of set X and $\bigcap$ is the intersection operator returning the overlap of two bounding boxes. $NoFrames(X)$ is the number of frames in set X. By varying these constraints $t_{sr}, t_{ps}, t_{rt}, t_{pt}$ while keeping the other ones fixed at a very low value ($\varepsilon = 0.1$), we can have four sets of precision-recall curves. And recall and precision are combined into the traditional F-score as $F = \frac{2\times Precision\times Recall}{Precision+Recall}$. Denoting by the F-Score $F(,,,)$ depending on the quality constraints, we get:

$$I_{sr} = \frac{1}{N}F(t_{sr}, \varepsilon, \varepsilon, \varepsilon) \quad (14)$$
$$I_{ps} = \frac{1}{N}F(\varepsilon, t_{ps}, \varepsilon, \varepsilon) \quad (15)$$
$$I_{rt} = \frac{1}{N}F(\varepsilon, \varepsilon, t_{rt}, \varepsilon) \quad (16)$$
$$I_{pt} = \frac{1}{N}F(\varepsilon, \varepsilon, \varepsilon, t_{pt}) \quad (17)$$

where N is value specifying the number of samples for the numerical integration. The final performance indicator is the mean over these four values, which is denoted as I_{all}. For more details of the evaluation metric, please visit the ICPR HARL competition website: http://liris.cnrs.fr/harl2012/evaluation.html. Note that the groundtruth annotations were performed by ourselves as they are not provided by the organizer of the HARL competition.

We first show that using the depth induced constraints introduced in Subsection III-B, human subject/object detection can be made more accurate. To demonstrate this, we randomly choose 1000 human subject detection results with groundtruth manual labels by directly applying HOG-SVM detector without depth constraints from the testing video sequences. We then set the detection scores for those samples that violate any of the two depth constraints defined in Subsection III-B to zero. The comparison of the ROC curves with/without depth constraints is illustrate in Figure 5. We can note that using the depth constraints, large portion of the false detections are removed. Example of the human subject tracking result is shown in Figure 6. The upper two rows show the bounding boxes of the detections before tracking and the lower two rows show after temporal matching and tracking, spurious detections are removed.

Second, we show that modeling $3D$ spatial temporal contextual information improves the action detection performance especially for actions that involves interactions. We demonstrate this capability by taking the action *DI: discussion of two people* as an example. For recognizing and localizing this action, we applying both the $3D$ spatial temporal contextual attributes introduced in Subsection III-C

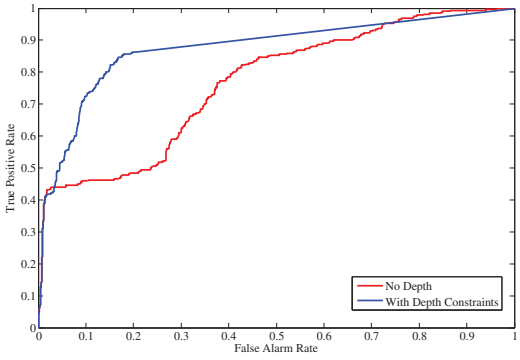

Fig. 5. ROC curves of the human subject detection results. Red: without the depth constraints; Blue: using the depth constraints.

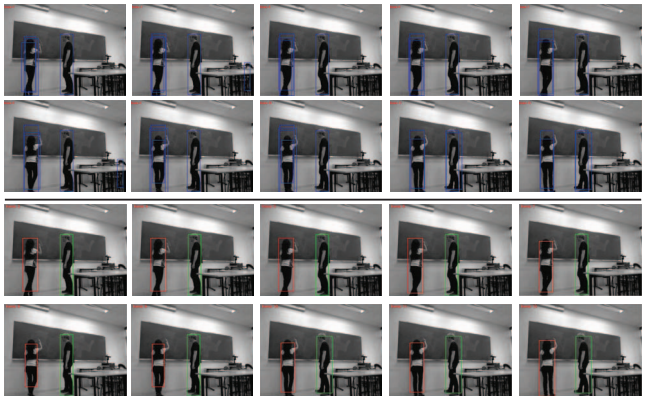

Fig. 6. Example of the tracking results. Upper two rows: before tracking. Lower two rows: after tracking. Red and green boxes indicate two tracked sequences. For better view, please zoom in the original pdf.

and its counterpart in $2D$, *i.e.*, only the $2D$ position, distance and velocity are considered using the grayscale image only. The prescion-recall curves and the corresponding F-score values are illustrated in Figure 7. We can note that using the $3D$ spatial temporal contextual information the false alarm rate can be reduced and the detection performance is significantly boosted.

Figure 8 shows the precision-recall curves for the results of all the 10 action classes. We also compare the performance with the state-of-the-art action detection and localization method [17] in Table I. We can note that the proposed method outperforms the previous art. Several example frames of the action localization results are given in Figure 9, where we can see that the localization is precise. Note that different instances of actions have large scale variations. Also, multiple actions can be detected simultaneously.

V. Conclusion and Future Work

In this paper, we present an action detection framework by utilizing the depth-induced contextual information from multiple stages of video processing for action representation. Experimental results demonstrate the effectiveness of the proposed scheme of fusing depth and grayscale images for three dimensional spatial and temporal interaction contex-

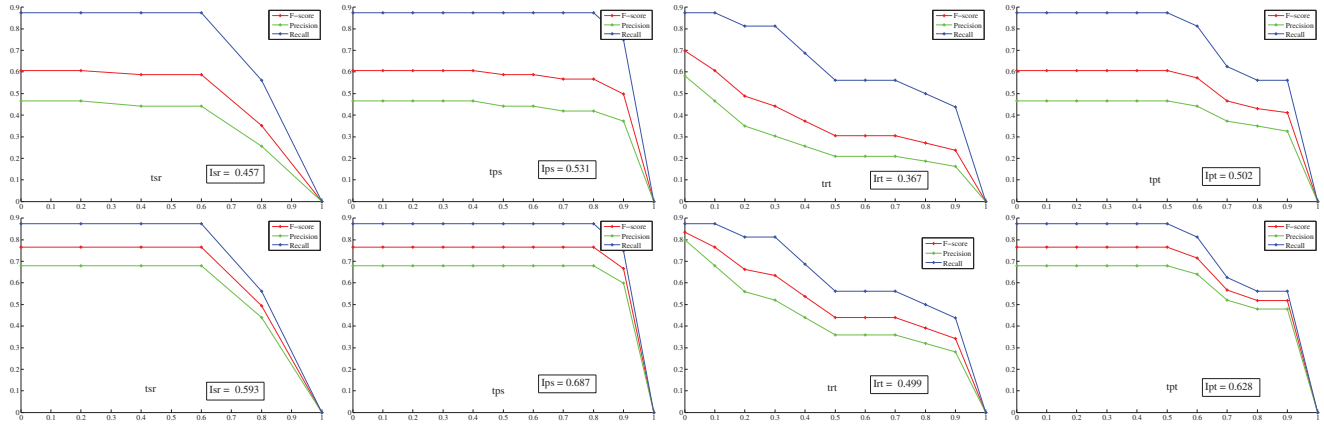

Fig. 7. For action *DI: Discussion of two people*: precision-recall curves by varying the parameter t_{sr}, t_{ps}, t_{rt}, t_{pt} and the corresponding I_{sr}, I_{ps}, I_{rt}, I_{pt} values. The upper row results are obtained without the depth image (apply $2D$ spatial temporal contextual attributes). The bottom row results are obtained with the depth image (apply $3D$ spatial temporal contextual attributes). Note that the I_{all} scores given by $2D$ and $3D$ methods are 0.464 and 0.602, respectively.

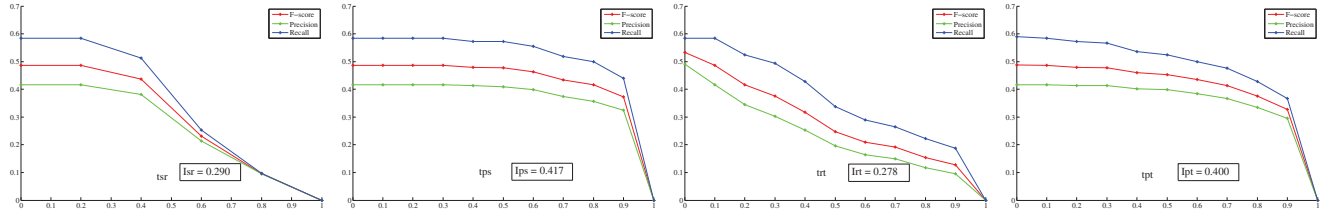

Fig. 8. For all action classes: precision-recall curves by varying the parameter t_{sr}, t_{ps}, t_{rt}, t_{pt} and the corresponding I_{sr}, I_{ps}, I_{rt}, I_{pt} values. Note that the I_{all} score is 0.346.

TABLE I

COMPARISONS OF PERFORMANCE (F-SCORES FOR ACTION RECOGNITION AND LOCALIZATION) WITH THE STATE-OF-THE-ART METHODS.

Measure	I_{sr}	I_{ps}	I_{rt}	I_{pt}	I_{all}
Yuan et. al [17]	0.214	0.367	0.225	0.358	0.291
Ours	0.290	0.417	0.278	0.400	0.346

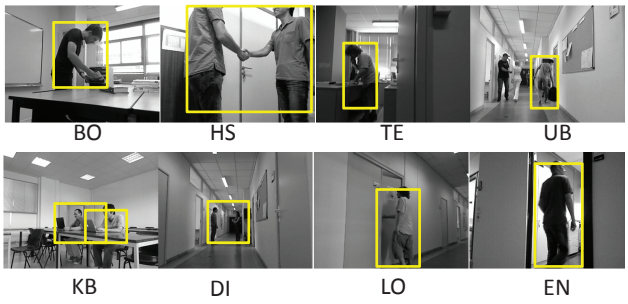

Fig. 9. Example frames of the action recognition and localization results.

tual modeling. We will investigate combining the proposed framework with depth based local motion features, *e.g.*, $3D$ joints' movement, for more accurate action recognition and localization.

VI. ACKNOWLEDGMENT

This study was supported by a research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapores Agency for Science, Technology and Research (A*STAR). This work was performed when Z. Liang was a research intern at the Advanced Digital Sciences Center, Singapore.

REFERENCES

[1] A. Bobick and J. Davis. The representation and recognition of action using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[6] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[7] K. Guo, P. Ishwar, and J. Konrad. Action recognition in video by sparse representation on covariance manifolds of silhouette tunnels. In *IEEE International Conference on Pattern Recognition*, 2010.

[8] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d gradients. In *The British Machine Vision Conference*, 2008.

[9] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in Neural Information Processing Systems*, pages 1216–1224, 2010.

[10] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision*, 2003.

[11] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPR Workshop on Human Communicative Behavior Analysis*, 2010.

[12] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *ICCV Workshops on Consumer Depth Cameras*, pages 1147–1153, 2011.

[13] B. Ni, S. Yan, and A. A. Kassim. Recognizing human group activities with localized causalities. In *CVPR*, pages 1470–1477, 2009.

[14] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3169–3176, 2011.

[15] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1290–1297, 2012.

[16] C. Wolf, J. Mille, L. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandrea, C. Bichot, C. Garcia, and B. Sankur. The liris human activities dataset and the icpr 2012 human activities recognition and localization competition. *Technical Report RR-LIRIS-2012-004, LIRIS Laboratory*, 2012.

[17] J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1728–1743, 2011.