Multilevel Depth and Image Fusion for Human Activity Detection

B. Ni, Y. Pei, P. Moulin, Fellow, IEEE, and S. Yan, Senior Member, IEEE

Abstract—Recognizing complex human activities usually requires the detection and modeling of individual visual features and the interactions between them. Current methods only rely on the visual features extracted from 2-D images, and therefore often lead to unreliable salient visual feature detection and inaccurate modeling of the interaction context between individual features.

In this paper, we show that these problems can be addressed by combining data from a conventional camera and a depth sensor (e.g., Microsoft Kinect). We propose a novel complex activity recognition and localization framework that effectively fuses information from both grayscale and depth image channels at multiple levels of the video processing pipeline. In the individual visual feature detection level, depth-based filters are applied to the detected human/object rectangles to remove false detections.

In the next level of interaction modeling, 3-D spatial and temporal contexts among human subjects or objects are extracted by integrating information from both grayscale and depth images. Depth information is also utilized to distinguish different types of interactions based on the detected human/object rectangles to remove false detections.

Finally, a latent structural model is developed to integrate the information from multiple levels of video processing (e.g., Microsoft Kinect). We propose a novel complex activity recognition and localization framework that effectively fuses information from both grayscale and depth image channels at multiple levels of the video processing pipeline. In the individual visual feature detection level, depth-based filters are applied to the detected human/object rectangles to remove false detections.

In the next level of interaction modeling, 3-D spatial and temporal contexts among human subjects or objects are extracted by integrating information from both grayscale and depth images. Depth information is also utilized to distinguish different types of interactions based on the detected human/object rectangles to remove false detections.

However, the performance of this individual visual feature + context model heavily relies on the quality of the visual feature detection and the accuracy of the contextual information encoding. On one hand, using state-of-the-art techniques in computer vision, very accurate detection of salient image or motion features such as human key poses, spatial-temporal interest points, motion trajectories, etc., is not achievable, due to the complex background. Current methods usually model the pairwise contextual information between individual visual features (e.g., two human key pose detections) based on their spatial and temporal displacement and relative velocity; however, as these quantities are measured in 2-D using conventional video cameras, ambiguity could arise and therefore the context encoding could be imprecise.

The recent emergence of depth sensors (e.g., Microsoft Kinect) has made it feasible and economically sound to capture in realtime not only color images, but also depth maps with good spatial resolution (e.g., 640 × 480) and accuracy. The data provide information about the 3-D structure of the scene as well as the 3-D motion of the subjects/objects in the scene. Therefore, ambiguity of the conventional camera, i.e., projection of the 3-D physical world onto the 2-D image
plane, could be circumvented. A comprehensive review on the applications of Kinect can be found in [13]. Intuitively, utilization of depth images can alleviate the above-mentioned problems (i.e., inaccurate individual visual feature detection and context modeling) and eventually benefit activity detection. First, the depth image provides layered information about the scene; therefore, detection of the salient foreground visual features, e.g., human key poses, could be more reliable, even in the presence of severe background clutters. Second, as the registered depth + image directly provides the 3-D coordinates of each scene point, accurate modeling of 3-D spatial and temporal relationship for representing various interaction becomes feasible. In addition, the depth map contains very rich 3-D structural information about the scene, therefore, the scene type can be represented more accurately by using depth map than by using conventional image. Note that the global scene contextual information is also an important cue for activity recognition.

Motivated by these observations, we propose a novel framework that effectively integrates depth information with the conventional grayscale image at different levels of the processing pipeline including: 1) individual feature extraction; 2) pairwise contextual information encoding, and 3) global scene representation. We show that using depth information in these processing stages can boost the performance of complex activity recognition and localization (i.e., activities contain various human/human, human/object, or human/surrounding interactions). More specifically, in the individual visual feature extraction step, the depth image is combined with the grayscale image for more robust human detection. During the pairwise context modeling step, 3-D spatial and temporal human/human, human/object, or human/surrounding relationships are directly obtained based on the registered grayscale and depth image. Also, depth information is utilized to extract the 3-D structure features for representing and classifying different types of indoor scenes. Finally, we integrate these feature representations from different levels using a latent structural support vector machine (SVM) model to form a discriminative activity detection framework. Extensive experiments on two activity recognition benchmarks (one with depth information) and a challenging grayscale + depth human activity dataset demonstrate the effectiveness of the proposed multilevel depth and image fusion scheme and the activity detection accuracy improvement over the previous art. The motivation for this paper is illustrated in Fig. 1.

The rest of this paper is organized as follows. First, we discuss related work in Section II. Section III presents the proposed multilevel depth and image fusion framework for activity recognition and localization. Extensive experimental results on two activity recognition benchmarks (one with depth information) and a challenging grayscale + depth human activity dataset are given in Section IV. Section V concludes the paper with discussions on future work.

II. RELATED WORK

Li et al. [14] 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 3-D 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. [15] developed the two color-depth fusion schemes for feature representation from the most representative feature representation methods in human action recognition. They first extended the STIFs into a depth-layered multichannel representation; then, they augmented the motion history images (MHIs) with the two depth-change-induced motion history channels. Superior performance is obtained by fusing color and depth information for human daily activity recognition. Liu and Shao [16] presented a learning method for selecting discriminative spatiotemporal features from RGBD sensor for video analysis.

It is generally agreed that knowing the 3-D joint position is helpful for action recognition. Based on the depth data and the estimated 3-D joint positions, Yuan et al. [17] proposed a new translational invariant local occupancy pattern feature, associated with each 3-D joint as the depth appearance of this 3-D 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 was 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. Sung et al. [18] detected and recognized human activities for the purpose of making personal assistant robots useful in performing assistive tasks. They used an RGBD sensor (Microsoft Kinect) as the input sensor, and presented a hierarchical maximum entropy Markov model (MEMM)-based learning algorithms to infer the activities. The features they used include the 3-D joint orientation matrix, hand position, and the motion information of the joints, obtained from the depth images. To achieve view-point invariance, the measurements are with respect to the center of torso.

Note that our work is not a competitor to these previous works. Rather, our proposed framework can be regarded as...
complementary to them. While the previous methods mainly work on how to extract low-level visual features and representation, e.g., the movement of the 3-D joints [17] or 3-D human pose and shape [14] using the depth image, our work focuses on effective integration of the depth information into multiple stages of the activity recognition processing pipeline. These stages include individual visual feature extraction, pairwise contextual encoding, and global scene representation. Our purpose is to improve the accuracy and robustness of these intermediary processing steps and to eventually boost the final activity detection performance. Also, we believe combining our proposed framework with previously developed depth-based 3-D motion or shape features can further boost activity detection performance.

III. METHODOLOGY

A. Motivation

Our approach is motivated by the observation that depth information can be well utilized to enhance the quality of representations in three stages of the action recognition pipeline, including: 1) human-centric action features (i.e., key poses); 2) human-to-human (object) interaction contextual features; and 3) global scene contextual features. These stages of feature representations have been widely used in the action recognition literature; however, previous works only utilize 2-D images/videos to obtain these features and often result in imprecise representations since real actions occur in 3-D. Therefore, our contribution is to enhance the feature extraction and representation in all three stages by using depth images. In particular, depth information is utilized to:

1) improve the accuracy of human key pose detection since the depth image provides good foreground/background segmentation cues;
2) measure human-to-human (object) interaction contexts including relative distance, velocity, and temporal ordering directly in 3-D, as this can remove the measurement ambiguity by using 2-D images only;
3) directly measure the 3-D scene structure, as it helps to improve the action recognition accuracy for these scene-dependent actions.

B. Overview of the Multilevel Depth and Image Fusion Scheme

Our basic idea is to integrate the depth information into multilevel processing pipeline for detecting human activities that involve human/human, human/object, or human/surrounding interactions. Fig. 2 illustrates our proposed processing pipeline for activity recognition and localization. Depth information at various processing stages is utilized and integrated in the following way. In the visual feature extraction step, we detect human key pose and object of interest in every input frame of the grayscale image sequence and the corresponding depth maps provide further constraints to filter out false detections. These human key pose and object detections are afterward spatially and temporally matched throughout frames into tracklets by applying the motion constraints in both grayscale and depth channel. As a byproduct, invalid detections without sufficient temporal durations are further filtered out at this stage. In the next stage, we model the 3-D spatiotemporal interaction/contextual attributes using combined grayscale and depth information. In the third stage, depth information is utilized for classifying the indoor scenes into different scene categories. Finally, the obtained spatial–temporal interaction attributes, key pose attributes of the tracklets and the scene classification results are integrated by a latent structural SVM for discriminatively recognizing and localizing activities.

C. Depth Constrained Human/Object Detection and Tracking

The first step in action recognition is to extract meaningful and representative visual features to measure the motion, shape, pose, etc., of the human subject(s) performing a certain activity. 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 paper, 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 variations in execution styles of the same action.

Human key poses are represented by a histogram of oriented gradient (HOG) features [19] extracted from grayscale image. In this paper, we use 8 × 8 blocks for calculating the gradient histogram, as in [20]. 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 the 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 a linear SVM detector based on HOG features [19] (HOG-SVM) using the training samples associated with each cluster (group, or key pose prototype). Similarly, we also detect objects of interest using the above method. Each detection instance \( x \) is assigned to a key pose type \( k = 1, 2, \ldots, K \), i.e., \( K \) key poses, via nearest cluster center assignment.

Using the grayscale image alone for human key pose or object detection often results in high false alarm rate, due to the background clutter and individual variations. Depth-based constraints can be used to effectively remove false detections. In this paper, we utilize two heuristic depth-based constraints.
1) The first constraint is that the area-to-median-depth ratio for a human subject should be within a certain range. For human subject detection \( x \), denote by \( \text{Area}(x) \) and \( d_m(x) \), the area and median depth, respectively. Then the constraint is given by:

\[
r_l \leq r(x) = \frac{\text{Area}(x)}{d_m(x)} \leq r_u
\]

where \( r(x) \) denotes the area to median depth ratio and \( r_l \) and \( r_u \) denotes the lower and upper bounds, respectively.

2) The second constraint is that the median human body depth value should be smaller (i.e., closer to the camera center) than the depth values surrounding the human body in the horizontal direction. To enforce this constraint, we denote \( d_m(lh) \) and \( d_m(rh) \) as the median depth value of the image stripe located on the left and right side of the human detection bounding box, respectively. The constraint can be expressed as:

\[
d_m(lh) > d_m(x) > d_m(rh). \tag{2}
\]

These two constraints are illustrated in Fig. 3. All the parameters can be estimated from the training data. In the experiment, we note that many false detections can be filtered out using these two depth-based constraints.

Once the candidate per frame subject/object detections are obtained, we temporally track them into human or object sequences named tracklets. The tracklet extraction process is based on pairwise detection matching over consecutive frames. The tracking process is as follows. First, once, we detect a new human key pose or object at some frame, we start from this detection (suppose the start frame number is \( i \)) and search its matched instance from all the detections obtained in the next frame \( i+1 \), i.e., we establish all the key pose or object detection matches between frames \( i \) and \( i+1 \). The matching score between a detection \( x \) in frame \( i \) and a detection \( y \) in frame \( i+1 \) is denoted as \( \text{dist}(x, y) \). Two detections are considered as matched if their \( \text{dist}(x, y) \) is smaller than some threshold value \( t_{\text{ms}} \). Also, we impose that for any detection in frame \( i \), there can be maximally one candidate match in frame \( i+1 \). This matching process is continued over frames until a stopping criterion is satisfied. The stopping criterion is defined as: no match is found in the subsequent three frames. The motivation for this stopping criterion is that due to occlusion or inaccuracy of HOG detector, there exist random missed detections at some frame. Therefore, checking three subsequent frames for a valid match instead of a single frame improves tracking robustness. Each matched detection sequence then forms a motion trajectory of the key poses or objects, which is a tracklet. Missed detections within a tracklet are linearly interpolated. Second, to remove possible noisy tracklets, we restrict the length \( L \) of any valid tracklet to be between \( L_{\text{min}} \) and \( L_{\text{max}} \). In this paper, we set \( L_{\text{min}} = 5 \) and \( L_{\text{max}} = 200 \), which correspond to 0.2–8 in duration.

To match two detections \( x \) and \( y \) in consecutive frames, we calculate the matching score, that is, the weighted combination of the 3-D spatial displacement between consecutive frames, the difference of detection bounding boxes and the appearance similarity as:

\[
d_{\text{ms}}(x, y) = a \parallel p(x) - p(y) \parallel^2 + \beta \parallel \text{Area}(x) - \text{Area}(y) \parallel^2 + \gamma \parallel \text{Hist}(x) - \text{Hist}(y) \parallel^2 \tag{3}
\]

where \( p(x) \) denotes the 3-D coordinate of the center of detection \( x \), i.e., \( p(x) = (x, y, z) \) and \( \text{Hist}(x) \) denotes the normalized grayscale histogram of the detection \( x \). The first term measures the distance proximity; the second term measures the normalized size difference of the detections; and the last term measures the appearance similarity. The weighting parameters \( a, \beta, \gamma \) and the distance threshold value \( t_{\text{ms}} \) are set by grid search and twofold cross validation based on the training data. In our experiment, we set \( a = 0.4, \beta = 10, \gamma = 20 \), respectively.

As mentioned previously, 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:

\[
m(k, j) = \frac{\left| \{ x : x \in \text{Pose}(k), x \in \text{Action}(j) \} \right|}{\left| \{ x : x \in \text{Pose}(k) \} \right|} \tag{4}
\]

Here, \( x \in \text{Pose}(k) \) means the detection \( x \) belongs to key pose type \( k \); \( x \in \text{Action}(j) \) means the detection \( x \) is contained in some instance of action class \( j \). \( \left| \{ \} \right| \) 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 \( m(k) = [m(k, 1), m(k, 2), \ldots, m(k, C)]^T \). Here, \( k = 1, 2, \ldots, K \), which indicates its action class association probability. For a tracklet \( T \), we then calculate its class probability \( f_k(T) \) by averaging the detections over the whole tracklet as:

\[
f_k(T) = \frac{1}{|T|} \sum_{x \in T} m(k(x)). \tag{5}
\]

Here, \( |T| \) denotes the number of elements in the tracklet \( T \), and \( k(x) \) denotes the type of key pose for detection \( x \). We refer \( f_k(T) \) as the single tracklet-based attribute for tracklet \( T \) in our integrated formulation for activity detection that will be introduced in Section III-D.
D. 3-D Spatial–Temporal Contextual Encoding

It is widely recognized that modeling spatial and temporal contextual information is important for representing human/human, human/object, or human/surrounding interactions [11]. Current methods mostly utilize conventional 2-D images, therefore, the contextual information such as relative distance, relative speed, etc., can only be measured in 2-D, and therefore inaccurate measurement and ambiguity exists due to perspective projection. When a depth image is available, spatiotemporal contextual information can be utilized to solve this problem easily. Using the depth image solves this problem easily. The advantage of modeling 3-D spatiotemporal contextual interaction between two objects is obvious. For example, the action discussion between two people requires understanding their relative positions and speeds. Using the depth image solves this problem easily. More specifically, we define the following 3-D spatiotemporal contextual interaction attributes between two tracklets \(T_1\) and \(T_2\):

1) Relative 3-D distance: The relative distance between two temporally overlapping tracklets over their common frames as

\[
L(T_1, T_2) = \frac{1}{|t_o(T_1, T_2) - t_e(T_1, T_2) + 1|} \times \sum_{i \in (t_o(T_1), t_e(T_1), t_o(T_2), t_e(T_2))} (\mathbf{p}(x_i^T) - \mathbf{p}(x_i^O)).
\]

Here, \(t_o(T)\) (i.e., stand for time for common start) and \(t_e(T)\) (i.e., stand for time for common end) mean the start and end frame number of the temporally overlapping portion of tracklet \(T_1\) and \(T_2\). If we denote \(t(T)\) as the start and end frame numbers of tracklet \(T\), then \(t_o(T_1, T_2)\) and \(t_e(T_1, T_2)\) can be computed as \(t_o(T_1, T_2) = \max(t_o(T_1), t_o(T_2))\) and \(t_e(T_1, T_2) = \min(t_e(T_1), t_e(T_2))\), respectively. \(\mathbf{x}_i^T\) denotes the human subject/object detection in \(i\)-th frame (with respect to the starting frame: \(t_o(T_1), T_2))\). Relative distance is useful for identifying some types of interactions such as discussion between two people, when the 3-D displacement between two human subjects has high discriminative capability. We note that \(L_T\) is a 3-D real vector.

2) Relative 3-D velocity: Similarly to previous works [21], [22], we define the mean relative velocity between two overlapping tracklets over their common frames as the following 3-D vector:

\[
L_v(T_1, T_2) = \frac{1}{|t_o(T_1, T_2) - t_e(T_1, T_2) + 1|} \times \sum_{i \in (t_o(T_1), t_e(T_1), t_o(T_2), t_e(T_2))} (\mathbf{v}(x_i^T) - \mathbf{v}(x_i^O)).
\]

Here, \(\mathbf{v}(x_i)\) denotes the 3-D velocity of the detection \(x\) in consecutive frames \(i\) and \(i+1\), i.e., \(\mathbf{v}(x_i) = \mathbf{p}(x_{i+1}) - \mathbf{p}(x_i)\). Relative 3-D 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 zero relative speed (i.e., unlock the door) and another phase of fixed relative speed (i.e., enter).

3) Relative temporal ordering: Temporal ordering relationship between two tracklets are also very important for distinguishing actions that 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)

\[
\min(t_o(T_1), t_o(T_2)) = \max(t_e(T_1), t_e(T_2)) + 1 - t_o(T).
\]

\(T_1\) is considered as preceding \(T_2\) when \(t_e(T_1) - t_o(T_2) < t_p\), and vice versa, \(t_c\) and \(t_p\) are the corresponding threshold values, which are estimated from the training data using grid search and twofold cross validation. We refer to \(L_o(T_1, T_2)\) as the temporal ordering attribute between tracklets \(T_1\) and \(T_2\). This is a 3-D binary vector.

E. 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-of-office scene (e.g., corridor). Similarly, the action “A person unlocks an office and then enters it” always occurs in a scene that contains a door. We note that when a depth image is available, we can use the 3-D geometric attributes for modeling the scene type. In particular, we can use the 3-D plane orientations. To model the geometric scene structure, we first transform the depth image into a normal map. The normal map is composed of values of each cell plane’s (patch) 3-D normal, which is a normalized vector that indicates the orientation of a plane. 3-D normals can be directly computed by fitting a plane to 3-D points sampled from a local image patch, that are assumed to be planar. For each pixel in the depth map, we use its 7 × 7 neighborhood pixels’ 3-D coordinates to fit a plane, and compute the plane’s normal. After calculating the normals, we project the 3-D plane directions onto the 2-D image plane. Finally, we represent the scene by: 1) the histogram of four major orientations (up, down, left, right), and 2) 2-D coordinates of the center of gravity of the four projected directions. These two types of feature vectors are concatenated to represent the scene, and linear SVMs [23] are trained to classify the scene. In this paper, we define 5 types of indoor scenes corresponding to typical scenarios of office environment. More scene types could be added for any new
where $T$ latent structural model, as in [24]. scene type attributes can be all modeled in a discriminative attributes, pairwise tracklet contextual attributes and the global $\phi_u$ on the pairwise contextual attributes of $(h_i,h_j)$, as $\phi_u(T(h_i),T(h_j),h_i,h_j) = [f_T(T(h_i),T(h_j))\phi_u, f_T(T(h_i),T(h_j))\phi_v]$. As mentioned before, $f$ is a feature vector that depends on the global scene attributes. $\theta$ denotes the set of parameters $\theta_s$, $\theta_w$, and $\theta_r$, and the bias $b$.

The goal is to learn the model parameters and construct the classifiers. Analogously to classical SVMs, we train from labeled examples $D$ that are partitioned into a positive sample set $P$ and a negative sample set $N$ by solving the following optimization problem:

$$
\min_{\theta} \frac{1}{2}\left\|\theta\right\|^2 + C \sum_{t=1}^{N} \xi_t,
$$

subject to

$$
\max \theta^T \phi(T^t, h) \geq 1 - \xi_t, \forall t \in P, \\
(1-\theta^T \phi(T^t, h)) \geq 1 - \xi_t, \forall t \in N,
$$

$$
\xi_t \geq 0, \forall t.
$$

IV. EXPERIMENTS

A. Activity Recognition on Cornell Activity Dataset

Our first experiment is performed on the Cornell RGB-D human activity dataset [18]. The video dataset is captured using the Kinect sensor, which produces 640×480 color-depth image sequences with human 3-D motion sequences, namely, each activity sample can be represented as a sequence of 3-D joint positions (or angles). The dataset consists of five scenarios: office, kitchen, bedroom, bathroom, and living room. Three to four common activities were identified for each location, giving a total of twelve unique activities collected from four subjects (with an additional neutral activity category). We use the leave-one-subject-out scheme; hence, subjects in the testing samples do not occur in the training samples. We compare the precision and recall values for various algorithms including (the results for the comparing algorithms are cited from [25]):

1) the hierarchical MEMM method proposed by Sung et al. [18];
2) the method based on object affordances proposed by Koppula et al. [26];
3) the eigensjoints method proposed by Yang and Tian [27];
4) our method. Note that this dataset do not contain human interactions, and therefore our terms based on 3-D spatiotemporal contextual representations are not applicable. Also, we do not apply the scene classification term and only the term for single tracklet-based attribute (i.e., first term in 9) in our method is used. For our
method, we report both results by using color image only and using color + depth images. The number of key pose types is $K = 25$.

The precision and recall values of the detection results from all comparing methods are summarized in Table I. Note that all compared methods except ours use the estimated 3-D joints information output by Kinect. This is a very strong feature for action representation; however, it cannot be reliably obtained in an uncontrolled and realistic environment, which limits the applicability of these methods. In contrast, our method achieves comparable results without using the 3-D joints data. We also note that using depth images improves action detection performance.

### B. Activity Recognition on Collective Activity Dataset

To evaluate our method’s capability in contextual information modeling, we test our algorithm on the collective activity dataset [28]. The dataset contains over 40 short video clips of crossing, waiting, queueing, walking, and talking action categories. The videos are $640 \times 480$ pixels in size and were recorded using a consumer hand held camera. Every 10th frame of all video sequence was manually labeled with pose, activity, and bounding box information. Following [28], a leave-one-out scheme was used to assess performance. We compare classification accuracies achieved by the video codewords method proposed in [29] and the spatiotemporal relationship method proposed in [28]. Since the evaluation settings are same for all methods, we directly cite the corresponding accuracy values reported in [28]. Since the evaluation settings are same for all methods, we directly cite the corresponding accuracy values reported in [28]. Note that this dataset does not contain depth images, therefore, all the contextual terms in our method are measured in 2-D. Also, we do not apply the third term of 9 corresponding to scene contextual attribute. For this dataset, we use eight key pose types ($K = 8$) corresponding to: right, front-right, front, front-left, left, back-left, back, and back-right human poses. Table II compares average detection accuracies, and we see that our method outperforms the others. This demonstrates our algorithm’s capability in encoding contextual information.

### C. Activity Recognition and Localization on ICPR 2012 HARL Contest Dataset

We use the HARL 2012 competition dataset [30] for an action recognition and localization experiment. The HARL 2012 competition focused on 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 was not only to classify activities, but also to detect and localize them. The dataset was 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 resolution of both grayscale and depth image are $640 \times 480$ pixels. The dataset contains grayscale/depth videos (D1) showing people performing various activities taken from daily life (discussing, making telephone calls, giving an item, etc.) The dataset is fully annotated, and the annotation not only contains information about the action class but also about its spatial and temporal positions in the video (bounding boxes). The dataset consists of 10 classes. Each class can be a normal activity, a human–human interaction, a human–object interaction, or a combination of the latter two types. These actions include the following:

1. DI: discussion of two or several people;
2. GI: a person gives an item to a second person;
3. BO: an item is picked up or put down;
4. EN: a person enters or leaves an office;
5. ET: a person tries to enter an office unsuccessfully;
6. LO: a person unlocks an office and then enters it;
7. UB: a person leaves baggage unattended;
8. HS: handshaking of two people;
9. KB: a person types on a keyboard; and
10. TE: a person talks on a telephone.

In total, the training set has 305 action samples and the testing set has 156 samples. On this dataset, we cluster the training human/object samples into 28 types of key poses.
Here, the best match (BM) is defined by

can only take the values one or zero:

d. In a similar way, we denote by \(d(g)\) the set of bounding boxes of the ground truth action \(g\) restricted to uniquely the frames that are also part of ground truth action \(d\). Then, \(IM(g, d) = 1\), when the following conditions are all satisfied [otherwise \(IM(g, d) = 0\), i.e., \(BM(\ldots)\) can only take the values one or zero]:

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 the two bounding boxes. NoFrames(\(X)\) is the number of frames in set \(X\). By varying these constraints \(I_{ps}, I_{pt}, I_{sr}, I_{rt}\) while keeping the other ones fixed at a very low value (\(\epsilon = 0.1\)), we obtain four sets of precision-recall curves. Recall and precision are combined into the traditional F-score as

\[ F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

The F-score is defined by \((\ldots)\) to make explicit its dependency on the quality constraints. We obtain

\[ I_{ps} = \frac{1}{N} F(t_{ps}, \epsilon, \epsilon, \epsilon) \]

\[ I_{pt} = \frac{1}{N} F(t_{pt}, t_{ps}, \epsilon, \epsilon) \]

\[ I_{sr} = \frac{1}{N} F(\epsilon, \epsilon, t_{sr}, \epsilon) \]

\[ I_{rt} = \frac{1}{N} F(\epsilon, \epsilon, \epsilon, t_{rt}) \]

where \(N\) is the value specifying the number of samples for numerical integration. The final performance indicator is the mean over these four values and is denoted by \(I_{F}\).

For more details of the evaluation metric, please visit the ICPR HARL competition website: http://liris.cnrs.fr/harl2012/evaluation.html.

We first show that using the depth-induced constraints introduced in Subsection III-C, 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 the HOG-SVM detector without depth constraints from the testing video sequences. We then set to zero the detection scores for those samples that violate any of the two depth constraints defined in Subsection III-C. The comparison of the ROC curves with and without depth constraints is illustrated in Fig. 5. We note that using the depth constraints, large portion of the false detections is removed. Two examples of human subject tracking results are shown in Fig. 6. For each example, the upper row shows the bounding boxes of the detections before tracking and the lower row shows after temporal matching and tracking, spurious detections are removed. Some removed spurious detections are highlighted by yellow boxes.

![Fig. 6. Example of the tracking results. Rows 1 and 3: before tracking. Rows 2 and 4: after tracking. Blue boxes indicated per frame detections (without temporal matching). Red and green boxes indicate two tracked sequences. Yellow boxes highlight some noisy detections.](image-url)
Second, we show that modeling 3-D spatiotemporal contextual information improves the action detection performance especially for actions that involves interactions. We demonstrate this capability by taking the action “DI: discussion between two people” as an example. For recognizing and localizing this action, we apply both the 3-D spatiotemporal contextual attributes introduced in Subsection III-D and their counterpart, i.e., only the 2-D position, distance and velocity are considered using the grayscale image only. The precision-recall curves and the corresponding F-score values are shown in Fig. 7. We note that using the 3-D spatiotemporal contextual information, the false alarm rate is reduced and the detection performance is significantly boosted.

In Fig. 8, we show example frames of the directional maps (calculated from the depth maps) of five representative scenes, their corresponding centers of gravity for different directions marked by circles, and the corresponding directional histograms. We note that different scenes have distinctive patterns of directional histograms and spatial distributions of different directions. The information obtained from the depth maps is very useful for scene classification.

### Table 1

<table>
<thead>
<tr>
<th>Action</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>10th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 1st: Left, 2nd: Right, 3rd: Up, 4th: Down, 5th: Left, 6th: Right, 7th: Up, 8th: Down, 9th: Left, 10th: Right
Fig. 11. Example frames of the activity recognition and localization results. Last three examples highlighted by red bounding box are failure cases.

<table>
<thead>
<tr>
<th>Measure</th>
<th>I_{PR}</th>
<th>I_{PF}</th>
<th>I_{RC}</th>
<th>I_{F1}</th>
<th>I_{ACC}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yuan et al. [31]</td>
<td>0.214</td>
<td>0.567</td>
<td>0.225</td>
<td>0.358</td>
<td>0.259</td>
</tr>
<tr>
<td>Wang et al. [24]</td>
<td>0.192</td>
<td>0.308</td>
<td>0.245</td>
<td>0.316</td>
<td>0.265</td>
</tr>
<tr>
<td>Ours</td>
<td>0.317</td>
<td>0.448</td>
<td>0.295</td>
<td>0.430</td>
<td>0.372</td>
</tr>
</tbody>
</table>

Fig. 9 shows the precision-recall curves for the results of all the 10 action classes. In addition, we also show in Fig. 10, the traditional confusion matrix that illustrates the pure classification performance by removing the spatiotemporal matching constraints defined above. Note that the confusion matrix ignores actions that have not been detected, and actions with no result. Therefore, unlike in classification tasks, the recognition rate (accuracy) cannot be determined from its diagonal. For this reason the confusion matrix must be accompanied by precision and recall values shown in Fig. 9. We also compare the performance with the state-of-the-art action detection and localization methods including: 1) the method proposed in [31], where STIP are extracted from representing actions and subvolume mutual information maximization is used to effectively search the best activity volume, i.e., localization, and 2) the part-based action recognition method proposed in [24]. For both methods, the best parameters are empirically tuned based on the training data. The comparisons are shown in Table III. We note that the proposed method greatly outperforms the previous art. Several example frames of the action localization results are given in Fig. 11. We have several observations: 1) for most examples, the activity spatial localization results are precise; 2) different instances of activities have very large scale variations; and 3) multiple activity instances can be detected simultaneously in the same frame. For the last three examples that are highlighted by red bounding box, the predicted activity class labels are incorrect. This is due to the fact that the current tracklet-based features cannot well distinguish the cases such as: stand together versus discussion, pass object versus shake hand. In the future, fine-grained local motion features should be developed to handle these cases. In addition, we also study the effect of the tracklet extraction parameter $t_{dist}$ on the final action detection performance. Fig. 12 shows different $I_{F1}$ scores obtained using different $t_{dist}$ values. We observe that $t_{dist} = 5.0$ is a reasonable
choice that we fix throughout all experiments in this paper. Larger \( t_{\text{fa}} \) values result in unreliable tracking, and therefore the final action detection performance is affected.

V. CONCLUSION

We have presented an activity detection framework that integrates multilevel depth information in the video processing pipeline to boost detection accuracy. Experimental results on two activity recognition benchmarks (one with depth information) and a challenging depth + grayscale activity dataset demonstrate the effectiveness of the proposed scheme of fusing depth and grayscale images for robust individual visual feature extraction, accurate 3-D spatial and temporal interaction contextual modeling, and the high-detection accuracy for complex activity and interaction.

REFERENCES


Jiang Su received the B.Eng. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2003, and the Ph.D. degree from the National University of Singapore, Singapore, in 2011. He is currently a Research Scientist with Advanced Digital Sciences Center, Singapore. His current research interests include the areas of computer vision, machine learning, and multimedia. He was with Microsoft Research Asia, Beijing, China, as a Research Intern in 2008. He was also a Software Engineer Intern with Google Inc., Mountain View, CA, USA, in 2010.

Dr. Ni was a recipient of the first prize in International Contest on Human Activity Recognition and Localization in conjunction with International Conference on Pattern Recognition in 2012.
Pierre Moulin (F’03) received the Doctoral degree from Washington University, St. Louis, MO, USA, in 1990. He joined Bell Communications Research, Morristown, NJ, USA, as a Research Scientist. In 1996, he joined the University of Illinois at Urbana-Champaign (UIUC), Urbana, IL, USA, where he is currently a Professor with the Department of Electrical and Computer Engineering, Research Professor with the Beckman Institute and the Coordinated Science Laboratory, and an Affiliate Professor with the Department of Statistics. His fields of professional interest include image and video processing, compression, statistical signal processing and modeling, media security, decision theory, and information theory.

Dr. Moulin has served on the editorial boards of the IEEE TRANSACTIONS ON INFORMATION THEORY and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is the Co-Founding Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He has served the IEEE in various other capacities and is currently a member of the IEEE Signal Processing Society Board of Governors. He was the recipient of the 1997 Career Award from the National Science Foundation and the IEEE Signal Processing Society 1997 Senior Best Paper Award. He is also a co-author (with Juan Liu) of a paper that received the IEEE Signal Processing Society 2002 Young Author Best Paper award. He was a 2003 Beckman Associate of UIUC’s Center for Advanced Study. He was a recipient of UIUC’s 2005 Sony Faculty Award, and a Plenary Speaker for ICASSP 2006.

Shuicheng Yan (M’06, SM’09) is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the Co-Founder of the Learning and Vision Research Group (http://www.lv-nus.org). He has authored/co-authored over 300 technical papers in a wide range of research topics, with Google Scholar citation > 9,300 times and H-index-42. His current research interests include computer vision, multimedia and machine learning.

Mr. Shuicheng is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (IEEE TCSVT) and ACM Transactions on Intelligent Systems and Technology (ACM TIST), and has been serving as the Guest Editor of the special issues for TMM and CVIU. He was a recipient of the Best Paper Awards from ACM MM’12 (demo); PCM’11, ACM MM’10, ICME’10 and ICMCS’09, the winner prizes of the classification task in PASCAL VOC from 2010 to 2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honorable mention prize of the detection task in PASCAL VOC’10, 2010 TCSVT Best Associate Editor Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.