# Adaptive Human-Centered Representation for Activity Recognition of Multiple Individuals from 3D Point Cloud Sequences

Hao Zhang[1], Christopher Reardon[2], Chi Zhang[2], and Lynne E. Parker[2]

***Abstract*— Activity recognition of multi-individuals (ARMI) within a group, which is essential to practical human-centered robotics applications such as childhood education, is a particularly challenging and previously not well studied problem. We present a novel adaptive human-centered (AdHuC) representation based on local spatio-temporal features (LST) to address ARMI in a sequence of 3D point clouds. Our human-centered detector constructs affiliation regions to associate LST features with humans by mining depth data and using a cascade of rejectors to localize humans in 3D space. Then, features are detected within each affiliation region, which avoids extracting irrelevant features from dynamic background clutter and addresses moving cameras on mobile robots. Our feature descriptor is able to adapt its support region to linear perspective view variations and encode multi-channel information (i.e., color and depth) to construct the final representation. Empirical studies validate that the AdHuC representation obtains promising performance on ARMI using an Meka humanoid robot to play multi-people Simon Says games. Experiments on benchmark datasets further demonstrate that our adaptive human-centered representation outperforms previous approaches for activity recognition from color-depth data.**

## I. Introduction

In this paper, we address the important but previously not well studied robot reasoning problem: mobile robot interpretation of behaviors of each individual in a group of humans. In particular, we are interested in the physical game playing scenario in childhood education applications. The objective of this work is to enable human-centered robotic systems to simultaneously interpret behaviors of multiple individuals in Simon Says games, as illustrated in Fig. 1, with the ultimate goal of deploying such a robotic system to simultaneously communicate and interact with multiple people in real-world human social environments.

The proposed approach also has significant potential for use in a large number of other real-world human-centered robotics applications, in which robots need the critical ability to understand behaviors of multiple individuals. For example, a robotic guard that observes multiple individuals should be able to recognize each person's movement to detect abnormal patterns; a service robot requires the capability of perceiving behaviors of each individual in a group to provide effectively for human needs; a self-driving robotic car needs to be able to distinguish each pedestrian's behaviors to better ensure safety; a medical robot needs to understand the motions of each patient in a group to evaluate treatment progress in physical therapy applications.

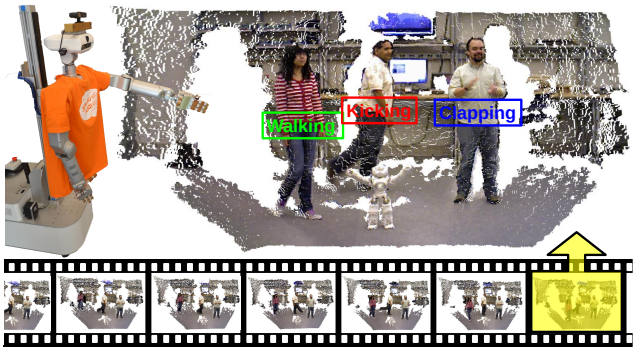



Fig. 1. A motivating example of ARMI: a Meka humanoid robot is used to play Simon Says games with a group of three individuals. The objective of ARMI is to recognize activities of each individual in a group, such as the walking activity performed by the female. Our solution is based on a new representation, which coherently and efficiently localizes people and extracts adaptive, human-centered LST features in $xyzt$ space. Our representation is able to (1) identify feature affiliations, i.e., which features are from whom, (2) avoid extracting irrelevant information from dynamic background and foreground obstacles, especially when a camera is moving with the robot, (3) address the false descriptor size problem through estimating the depth of a feature's support region and adapting its size to compensate for linear perspective view changes.

Although several approaches are discussed in the robotic perception literature to address the task of single-person action reasoning [1], [2], [3] and group action recognition [4], [5], [6], [7], interpreting the actions of each or a specific individual within a group has not been previously well investigated. We name this essential problem *activity recognition of multiple individuals* (ARMI), as depicted in the Simon Says game in Fig. 1. In real-world human social environments, including schools, hospitals, business, etc., a human-centered robot can experience complex, dynamic environments with multiple humans present. The capability of performing ARMI in these settings, within complicated scenes with camera motions, occlusions, background clutter, and illumination variations, is therefore extremely significant.

Among different approaches for representing humans and their behaviors [2], [4], [5], [6], [7], local spatio-temporal (LST) features are the most popular and promising representation. Human representation approaches based on LST features are generally invariant to geometric transformations; as a result, they are less affected by variations in scale, rotation and viewpoint. Because LST features are locally computed, they are inherently robust to occlusions. Use of orientation-

[1]Hao Zhang is with the Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401, USA. hzhang@mines.edu

[2]Christopher Reardon, Chi Zhang and Lynne E. Parker are with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA. {creardon, czhang24, leparker}@utk.edu

based descriptors provides the representation with additional robustness to illumination variations. Recently, the popularity of using affordable structured-light cameras to construct 3D robotic vision systems, in which human representations are developed and valuable depth information is encoded into LST features, continues to attract increasing attention from computer vision [2], [8] and robotics communities [3].

Despite their advantages, existing representations based on LST features have several shortcomings. First, because local features ignore global spatial structure information and lack affiliation information, they are incapable of identifying the activities of multiple persons in the same scene. Representing behaviors of each individual in a group is considerably more challenging than representing the group as a whole, which requires modeling *feature affiliation*, i.e., to which individual each feature is affiliated. Second, since representations based on LST features represent local variations, in complex scenes a large proportion of detected features often fall on the cluttered background, especially when the camera is moving, as is common with mobile robots. In this case, irrelevant features from backgrounds usually decrease the ability to represent human activities themselves [9]. Third, existing LST feature descriptors are generally not adaptive to linear perspective view variations, i.e., the size of the feature's support region does not adapt to the distance to the camera, which results in decreased feature description capability. For example, the values of local features extracted from the same point on a human can vary significantly when the human is positioned at different distances from the camera. We name this problem the *false descriptor size issue*.

In this paper, we address the problem of simultaneously recognizing activities of each individual in a group from 3D point cloud sequences, by proposing a novel representation based on *Ad*aptive *Hu*man-*C*entered (AdHuC) LST features that can address the above three shortcomings. Specifically, we construct affiliation regions of each human performing actions in $xyzt$ space (3D spatial and 1D temporal). Then, features are detected locally within an individual's affiliation region, which are assumed to be affiliated with the human. As a result, our human-centered feature detection technique explicitly models feature affiliation to solve the ARMI task and avoids detecting irrelevant local space-time features from background clutter. For feature description, we define a new depth measure to represent the true depth of a feature support region; we adaptively resize the support region based on this depth to compensate for linear perspective view changes, and we implement a novel normalized multi-channel descriptor to quantize our features. Our feature extraction approach is based on the depth of interest concept, which enables us to coherently localize humans, construct affiliation regions, and extract features to construct our AdHuC representation.

Our contributions are threefold. First, we propose a novel multi-channel feature detector to detect human-centered features from 3D point cloud sequences, which can represent activities of each individual in a group of humans and deal with background clutter and camera movements. Second, we introduce a new multi-channel descriptor that is able to compensate for the linear perspective view change and solve the false descriptor size problem. Third, we introduce a novel, coherent framework to simultaneously perform human localization and feature extraction, in order to construct the AdHuC representation to address ARMI at the feature level.

## II. Related Work

We review representations that can be adopted by robotic systems to represent people. We also review feature detectors and descriptors to construct human representations based on local features to recognize activities from 3D visual data.

### A. Human Representation in 3D Space

Several representations of humans in 3D space were proposed in the past few years. A naive 3D human representation approach is based on the human centroid trajectory in 3D space [10]. Although this representation provides a compact description of humans in a large space, it cannot be used to represent fine-grained human behaviors with limb motions.

Another category of human representations is based on human shape information, including the 3D human silhouette history [11]. Similar representations were also implemented based on 3D body-part models [12]. However, the robustness of these human representations is heavily restricted by the performance of human segmentation and body part tracking, which are also extremely challenging research problems due to background clutter, occlusion and camera motions.

In recent years, skeleton-based 3D human representation has received an increasing attention, since skeleton data are directly available from structured-light 3D sensors, such as the Kinect and PrimeSense. For example, [13] proposed a 3D representation based on the joint rotation matrix with respect to human torso. A representation based on actionlet ensemble was introduced in [14] to recognize activities from skeletal data. Other skeleton based 3D human representations were implemented using histograms of oriented displacements [15], covariance of 3D joints [16], etc. Since these representations completely rely on the skeleton data from color-depth cameras, they typically do not work in outdoor environments due to the sensing limitation of structured-light sensors. Also, skeleton data acquired from color-depth sensors can become inaccurate and very noisy when camera motions, background clutter, and occlusion are present.

### B. Local Spatio-Temporal Features

A most widely used human representation is based on local spatio-temporal features, which are typically extracted using two procedures: feature detection and description.

*1) Feature detection:* LST features are detected by capturing local texture and motion variations. Laptev et al. [17] detected LST features from color videos based on generalized Harris corner detectors with spatio-temporal Gaussian derivative filters. Dollar et al. [18] detected such features using separable filters in space and time dimensions from color videos. Recently, Zhang and Parker [3] extended [18] to detect features in color-depth videos. These methods extract LST features from the entire frame; as a result, they detect a

large portion of irrelevant features from background clutter and are incapable of distinguishing features from different individuals in a group.

Chakraborty et al. [9] proposed the selective LST feature, where interest points are extracted from the entire image and then features are pruned using surrounding suppression and space-time constraints. Our feature detector is inherently different from these feature selection methods; features irrelevant to humans are not detected (that is, no selection is performed), which significantly reduces the number of irrelevant features and thus decreases computational requirements, especially when the camera is in motion in robotics applications.

*2) Feature description:* Nearly all LST feature descriptors applied to represent human activities in videos are based on image gradients. Dollar et al. [18] concatenated image gradients within a fixed support region into a single feature vector. Zhang and Parker [3] extended [18] to describe multi-channel features with a fixed support region in $xyzt$ space. Scovanner et al. introduced the SIFT3D [19] descriptor to quantize gradients in space-time dimensions. Klaser et al. [1] introduced the HOG3D descriptor to describe $xyt$ gradients in a fixed support region. The support regions of these LST descriptors have a fixed, nonadaptive size and are not capable of handling the linear view perspective changes.

An approach was introduced in [2] to adapt support region size: detect features from entire frames and then assign each feature depth with the minimum depth value of the feature point within a time interval. This detect-assign approach suffers from the false descriptor size issue, because as most LST features are detected around the edges of moving body parts, a large proportion of local features fall outside of the blob of human pixels with incorrect depth values within the time interval. This could improperly treat these features as belonging to either the incorrect background or foreground objects, which would yield improper support region sizes. Inherently different from [2], we first analyze the depth to construct a feature affiliation region for each human, then detect features within each affiliation region. Since feature depth is constrained by affiliation regions, our descriptor is able to appropriately address the false descriptor size issue.

## III. Our Approach

Our objective of introducing the adaptive human-centered representation is to affiliate LST features with proper humans (i.e., human-centered) and adapt to linear perspective view changes (i.e., adaptive), in order to efficiently address ARMI in practical human-centered robotic applications.

### A. Mining Depth Information

We previously introduced an approach of analyzing depth information to discover depth of interest, which is a highly probable interval of human or object instances in the depth distribution of 3D point cloud sequences [20]. This concept is used as a foundation of our coherent framework for human localization and feature extraction. Each instance in a depth of interest is referred to as a candidate.

The input to our method is a sequence of 3D point clouds [21] or color-depth images acquired by a color-depth camera or computed from stereo cameras. When the camera operates on the same ground plane as humans (e.g., installed on mobile robots), ground and ceiling planes are typically viewable. Because points on the ground always connect candidates on the floor, it is important to eliminate this connection in order to robustly localize depths of interest that contain separate candidates of interest. Since a ceiling plane usually consists of a significant amount of irrelevant 3D points that gradually change depth, removing these points is desirable to increase processing efficiency, which is important for onboard robotic applications. To remove these planes, we apply the RANdom SAmple Consensus (RANSAC) approach [22], which is an iterative data-driven technique to estimate parameters of a mathematical model.

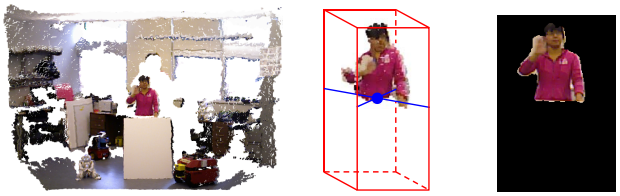
With ground and ceiling planes removed, a local maximum in the depth distribution is selected as a depth of interest: a depth interval centered at that maximum generally has a high probability to contain candidates that we are interested in. The correctness of our approach can be supported by the observation that, in 3D point clouds, any candidate contains a set or adjacent sets of points with a similar depth value. Since the 3D scene observed by a robot is typically dynamic and its underlying density form is unknown, we employ the non-parametric Parzen window method [23] to estimate the depth distribution and select candidates.

### B. Affiliation Region Construction

*Affiliation region* is defined in $xyzt$ space as a temporal sequence of cubes in 3D ($xyz$) space, such that each affiliation region contains *one and only one* individual with the same identity, which is denoted as $\mathcal{A}_h = \{x, y, z, t, s_x, s_y, s_z\}$, with cube center $(x, y, z)$ and size $(s_x, s_y, s_z)$ at time point $t$, and human identity $h$. The goal of introducing affiliation regions is to set constraints on locations where features can be extracted and to associate the features with the human in an affiliation region. To construct affiliation regions, humans are localized in $xyz$ space in each 3D point cloud frame, then locations of the same human across frames are associated.

Human localization is performed based on depth of interest. To preserve human candidates in each depth of interest, a cascade of rejectors is used to reject candidates that contain only non-human objects:

1) Height-based rejector: after a candidate's actual height is estimated, as illustrated in Fig. 2(b), it is rejected if its height is smaller than a min-height threshold (e.g., the NAO robot), or larger than a max-height threshold.
2) Size-based rejector: After estimating the actual size of a candidate, the candidate is rejected if its size is greater than a max-size threshold. However, to allow for occlusion, we do not reject small-sized candidates.
3) Surface-normal-based rejector: This detector is applied to reject planes, such as walls and desk surfaces, which cannot be humans.
4) HOG-based rejector: This rejector is based on a linear SVM and the HOG features, as proposed by Dalal and



(a) (b) (c)

Fig. 2. Computation of the height and centroid of an occluded candidate. Fig. 2(a) shows a raw color-depth frame. The actual height of a candidate is defined as the distance between its highest point to the ground, as shown in Fig. 2(b). The candidate centroid is drawn with a blue dot in the center of the 3D cube in Fig. 2(b). When the candidate is projected to a color image of size $96 \times 64$, it is placed in the center of the image according to its real size, instead of the blob size, as shown in Fig. 2(c).

Triggs [24]. As illustrated in Fig. 2, by using the actual height rather than a candidate's blob height, we obtain a more reliable rejection result.

In our cascade, simple rejectors are first used to reject the majority of candidates before more complex approaches are applied, which can significantly increase detection accuracy while radically reducing computation cost.

To associate human localization results across frames, an efficient loose-tight association method is introduced. Loose association is based on localized human candidate positions: if the distance of a human candidate in the current frame to a human in the previous frame is smaller than a predetermined threshold, they are loosely matched. Then, tight association is performed to further match loosely-associated human localization results. We create a color-based appearance model for each localized human, which is learned and updated in an online fashion using an online AdaBoost algorithm as by Grabner et al. [25].

Our affiliation region construction has several advantages: (1) Our color-based appearance model is an accurate human representation, since the background is masked out in color images, as demonstrated in Fig. 2(c). (2) Since our human appearance model is updated online, it adapts to appearance changes caused by occlusions and body configuration variations. (3) Human localization, based on depth of interest and rejector cascade, avoids computationally expensive window scanning over the entire frame and is able to localize humans using a moving robotic vision system, which is critical for mobile robots with computational constraints.

### C. Human-Centered Multi-Channel Feature Detection

Given a sequence of color-depth frames containing depth $d(x, y, z, t) = z(x, y, t)$ and color $c(x, y, z, t)$ data in $xyzt$ space, and the affiliation regions $\{\mathcal{A}_1, \ldots, \mathcal{A}_H\}$ constructed for $H$ humans in the camera view, our goal is to detect multi-channel LST features that are affiliated with people, which are called multi-channel *human-centered features*. Different from previous feature detection methods that detect interest points from entire frames without extracting the affiliation information [2], [3], [6], [18], we detect our human-centered features within the affiliation region of each individual, and associate the extracted feature's affiliations with the human.

To incorporate spatio-temporal and color-depth information in $xyzt$ space, we implement a cascade of three filters: a pass-through filter to encode cues along depth ($z$) dimension, a Gaussian filter to encode cues in $xy$ space, and a Gabor filter to encode time ($t$) information; then, we fuse the color and depth cues. Formally, within the affiliation region $\mathcal{A}_h$ of individual $h$, we convert color into intensity $i(x, y, z, t)$ and compute a multi-channel saliency map by applying separable filters over depth $d(x, y, z, t)$ and color $i(x, y, z, t)$ channels in $\mathcal{A}_h$. Depth and intensity data are processed using the same procedure. First, the data are filtered in the 3D spatial space:

$$d_s(x, y, z, t) = \big(d(x, y, z, t) \circ f(z, t; \delta)\big) * p(x, y; \sigma) \quad (1)$$

where $*$ denotes convolution and $\circ$ denotes entry-wise matrix multiplication. A pass-through filter $f(z, t; \delta)$ with parameter $\delta$ is applied along the $z$ dimension:

$$f(z, t; \delta) = H(z + \delta) - H(z - \delta) \quad (2)$$

where $H(\cdot)$ denotes the Heaviside step function. A Gaussian filter $p(x, y, t; \sigma)$ is applied along the $xy$ spatial dimensions:

$$p(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (3)$$

where $\sigma$ controls the spatial scale along $x$ and $y$ dimensions. Then, a Gabor filter is used along the $t$ dimension:

$$d_{st}(x, y, z, t) = d_s(x, y, z, t) * g(t; \tau, \omega) \quad (4)$$

where the Gabor filter $g(t; \tau, \omega)$ with parameter $\tau$ satisfies:

$$g(t; \tau, \omega) = \frac{1}{\sqrt{2\pi}\tau} \cdot e^{-\frac{t^2}{2\tau^2}} \cdot e^{i(2\pi\omega t)} \quad (5)$$

We use $\omega = 0.6/\tau$ throughout the paper.

After processing intensity data, we use the same procedure to obtain $i_{ds}(x, y, z, t)$. Then, we compute the spatio-temporal multi-channel saliency map as:

$$R(x, y, z, t) = (1 - \alpha) \cdot i_{st}^2(x, y, z, t) + \alpha \cdot d_{st}^2(x, y, z, t) \quad (6)$$

where $\alpha$ is a mixture weight to balance between intensity and depth cues. The saliency map generally represents variations of textures, shapes and motions, since any region undergoing such variations induces responses.

Next, our human-centered LST features are detected as local maximums of $R$ on the surface $z = z(x, y, t)$ within $\mathcal{A}_h$ in $xyzt$ space, and each feature is affiliated with human $h$. Since $\mathcal{A}_h$ only contains a single individual, the detected features are affiliated with the human and are distinguishable from features belonging to other individuals. As a result, our human-centered features are able to address the ARMI task. In addition, because the region for detecting our features is bounded by $\mathcal{A}_h$ in a depth of interest, irrelevant features (e.g., from the background) are never detected. This characteristic is particularly impactful in mobile robotic applications, since it provides (1) an increased description power to accurately represent humans, (2) an improvement in feature detection efficiency, and (3) the ability to handle the moving camera challenge in human representation based on local features.

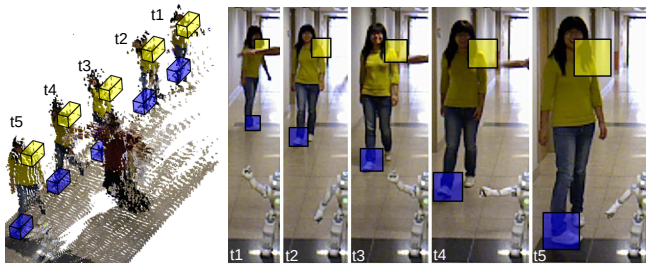



Fig. 3. Feature support regions that have the same size in 3D ($xyz$) physical space have different projected sizes when they are mapped onto 2D ($xy$) images, due to linear perspective view changes, as illustrated by the yellow and blue support regions. Accordingly, our adaptive multi-channel feature descriptor adapts support region sizes to their depth to compensate for this linear perspective view variation.

### D. Adaptive Multi-Channel Feature Description

Here we introduce a new multi-channel feature descriptor using a support region that is adaptive to changing linear perspective views, and thereby addresses the false descriptor size issue. For each LST feature point $(x, y, z, t, h)$, which falls in the affiliation region $\mathcal{A}_h = \{x_h, y_h, z_h, t_h, s_x, s_y, s_z\}$ and is detected with the 4D scale $(\sigma, \sigma, \delta, \tau)$ in $xyzt$ space, we extract a support region $\boldsymbol{S} = (x, y, z_s, t, \sigma_s, \delta_s, \tau_s, h)$ of size $(\sigma_s, \sigma_s, \delta_s, \tau_s)$ in $x$, $y$, $z$ and $t$ dimensions, respectively. To compensate for spatial linear perspective view changes, i.e., objects closer to a camera appearing larger, we propose adapting the spatial size of a support region to its depth. Estimating the true depth is a challenging task, since detected feature points can fall out of human blobs and consequently have incorrect depth values, resulting in the false descriptor size issue.

In order to address this issue, we propose a new approach to estimate the support region's depth, based on $\mathcal{A}_h$ that is computed using depths of interest. To this end, we formally define several important concepts and mathematically formulate our depth estimation statement as a proposition followed by a formal mathematical proof.

*Definition 1 (Depth affiliation indicator):* Given the affiliation region of an individual $\mathcal{A}_h = \{x_h, y_h, z_h, t_h, s_x, s_y, s_z\}$ and a depth value $z$, the *depth affiliation indicator* is defined as a function such that:

$$h(z) = \mathbb{1}(z \geq z_h - \frac{s_z}{2}) \cdot \mathbb{1}(z \leq z_h + \frac{s_z}{2}) \tag{7}$$

where $\mathbb{1}(\cdot)$ is the standard indicator function.

*Definition 2 (Support region's depth):* Given a local feature point $(x, y, z, t, h)$ detected using the scales $(\sigma, \sigma, \delta, \tau)$ in $\mathcal{A}_h = \{x_h, y_h, z_h, t_h, s_x, s_y, s_z\}$, the *depth* of the feature's support region $\boldsymbol{S}$ is defined by:

$$\begin{aligned} z_s(\boldsymbol{S}) = \frac{1}{\tau} \sum_{j=0}^{\tau-1} & z(x, y, t-j) \cdot h(z(x, y, t-j)) \\ & + z_h \cdot (1 - h(z(x, y, t-j))) \end{aligned} \tag{8}$$

*Proposition 1:* Given the affiliation region of an individual $\mathcal{A}_h$, for all feature points detected in $\mathcal{A}_h$, the true depth of their support regions satisfies $h(z_s(\boldsymbol{S})) = 1$.

*Proof:* Among a temporal sequence of $\tau$ depth values $z(x, y, t-j)$, $j = 0, \ldots, \tau-1$, assume $\tau_1 \in [0, \tau]$ out of $\tau$ depth values satisfy $h(z_i) = 1$, $i = 1, \ldots, \tau_1$; the remaining $\tau - \tau_1$ depth values satisfy $h(z_k) = 0$, $k = 1, \ldots, \tau - \tau_1$. Then, the support region's depth $z_s(\boldsymbol{S})$ satisfies:

$$\begin{aligned} z_s(\boldsymbol{S}) = & \frac{1}{\tau} \left( \sum_{i=1}^{\tau_1} z_i + \sum_{k=1}^{\tau-\tau_1} z_h \right) \\ \leq & \frac{1}{\tau} \left( \tau_1 \left( z_h + \frac{s_z}{2} \right) + (\tau - \tau_1) z_h \right) = \frac{s_z}{2} \frac{\tau_1}{\tau} + z_h \\ \leq & \frac{s_z}{2} + z_h \end{aligned} \tag{9}$$

Similarly, we can prove that $z_s(\boldsymbol{S})$ also satisfies:

$$\begin{aligned} z_s(\boldsymbol{S}) = & \frac{1}{\tau} \left( \sum_{i=1}^{\tau_1} z_i + \sum_{k=1}^{\tau-\tau_1} z_h \right) \\ \geq & \frac{1}{\tau} \left( \tau_1 \left( z_h - \frac{s_z}{2} \right) + (\tau - \tau_1) z_h \right) = z_h - \frac{s_z}{2} \frac{\tau_1}{\tau} \\ \geq & z_h - \frac{s_z}{2} \end{aligned} \tag{10}$$

In summary, $z_h - s_z/2 \leq z_s(\boldsymbol{S}) \leq z_h + s_z/2$. Therefore $h(z_s(\boldsymbol{S})) = 1$ ■

Proposition 1 shows that the location of the support region $\boldsymbol{S}$ is bounded by $\mathcal{A}_h$. Thus, $z_s(\boldsymbol{S})$ encodes the true depth of the support region $\boldsymbol{S}$ in $\mathcal{A}_h$, in general. Based on $z_s$, we adapt the spatial support region size as follows:

$$\sigma_s = \frac{\sigma_0 \sigma}{z_s}, \qquad \delta_s = \frac{\sigma_0 \delta}{z_s} \tag{11}$$

where $\sigma_0$ characterizes the support region's relative spatial size. Since its temporal size is not affected by spatial linear perspective view variations, we define $\tau_s = \tau_0 \tau$, where $\tau_0$ characterizes the relative temporal size. An example of our adaptive feature description is illustrated in Fig. 3.

We implement an extended HOG3D descriptor that slightly differs from the original [1] to incorporate multi-channel information and deal with adaptive supporting size. HOG3D approximates orientations of 3D gradients in a support region using a regular polyhedron with congruent regular polygon faces as bins. Tracing each gradient along its direction up to the intersection with a face identifies the bin index. Then, a feature is described as a histogram $\boldsymbol{h}$ that counts the number of gradients falling in the bins. Since the size of our feature's support region is adaptive, it can contain a different number of gradients; thus, histogram normalization is necessary. In addition, in order to incorporate information computed from both intensity and depth channels, we employ the standard practice of concatenation of the per-channel descriptors [3], [8], leading to our final descriptor:

$$\boldsymbol{h} = \left\{ \frac{\boldsymbol{h}_i}{M_i}, \frac{\boldsymbol{h}_d}{M_d} \right\} \tag{12}$$

where $\boldsymbol{h}_i$ is the histogram using $M_i$ intensity gradients, and $\boldsymbol{h}_d$ is the histogram based on $M_d$ depth gradients.

## IV. Implementation

For affiliation region construction, the width of depth of interest is set to 1.0 m; the min-height threshold is set to 0.4 m; the max-height threshold is set to 2.3 m; and the max-size threshold is set to 4.0 $m^2$. Our HOG-based rejector is modified from [24] and trained with the H3D dataset [26], using all positive and a subset of negative samples; the loose association threshold is set to be 0.5 m. For human-centered LST feature detection, we assign scale parameters $\sigma = 5$, $\delta = 0.25$ m, and $\tau = 3$. For adaptive multi-channel LST feature description, we assign parameter values $\sigma_0 = 8$ and $\tau_0 = 5$. When a color-depth camera is employed (e.g., Kinect), the depth value is in [0.5, 8.0] m. A standard feature pooling scheme [1], [2], [8], [27], [28] is applied for human activity recognition, which subdivides each support region into $N_x \times N_y \times N_t = 4 \times 4 \times 3$ cells.

Following the common practice [1], [2], [8], [28], human activity recognition is performed using a standard bag-of-features learning framework and a codebook is created by clustering 200,000 randomly sampled features using $k$-means into 1000 codewords. For classification, we use non-linear SVMs with $\chi^2$-kernels and the one-against-all approach [1], [8], [28]. The recognition method and our AdHuC representation are implemented using a mixture of Matlab and C++ in Robot Operating System (ROS) on a Linux machine with an i7 3.0G CPU and 16Gb memory.

## V. Experiments

To evaluate the performance of our AdHuC representation on activity recognition of multiple individuals, we conduct comprehensive experiments using a physical Meka robot in a Simon Says gaming scenario. In addition, to demonstrate our representation's impact, we compare our approach with methods in previous studies using publicly available benchmark datasets.

### A. ARMI in Multi-Human Simon Says Games

The goal of this task is to enable a Meka humanoid robot to simultaneously interpret behaviors of multiple individuals in a Simon Says game. Simon Says is a multi-player game, where one person plays "Simon" and issues instructions that should only be followed if prefaced with "Simon says".

*1) Experiment Settings:* The Meka M3 mobile humanoid robot is applied in the experiments, which is equipped with two 7-DOF elastic arms with 6-DOF force torque sensors, two 5-DOF hands, and a torso on a prismatic lift mounted on an omnidirectional base. In particular, its sensor head has two PrimeSense cameras (short and long range cameras), one Point Grey Flea3 high-speed wide-angle camera, and one Point Grey Bumblebee XB3 stereo camera, as illustrated in Fig. 4(a). The 3D robotic vision, based on structured light sensors (i.e., PrimeSense) and stereo cameras (Bumblebee XB3), allows the robot to acquire color-depth and 3D point cloud data in both indoor and outdoor environments.

To evaluate our new AdHuC representation's performance on ARMI, we collected a dataset in the Simon Says scenario. In the experiment, the Meka robot played "Simon" and issued six activities that human players need to follow, including bending, clapping, flapping, kicking, walking, and waving. Two or three subjects participated in each experiment, performing different activities. The subjects were free to perform activities with either hand/foot and stand at any depth in the camera view. A total number of 522 instances were collected using a long-range PrimeSense sensor with a $640 \times 480$ resolution at 30 Hz. Our Simon Says experiments were conducted in a complex, realistic indoor environment with the challenges of dynamic background (TV and monitor screens), human-shape dynamic foreground obstacles (Nao robot), partial occlusions, illumination variations, and camera motions with the Meka robot, as shown in Fig. 4(b).

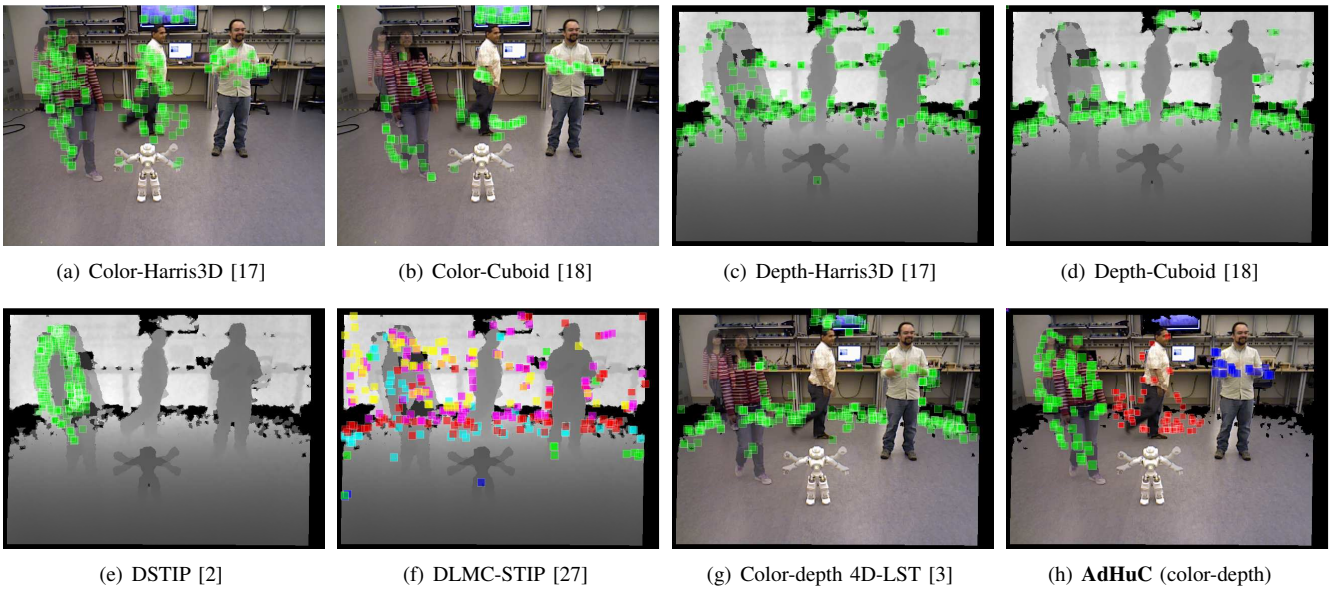
(a) Meka robot and its sensor head (b) Observed 3D scene

Fig. 4. 3D scene observed by structured light sensor on an Meka humanoid robot in Simon Says games. In the game, Meka issues an instruction "Simon says: wave your hand"; the male player correctly follows this instruction, but the female player fails to follow it.

For evaluation purposes, the following all-in-one setup is applied: we divide all data instances into 50% training and 50% testing, both containing activities from all game participants. Recognition performance is evaluated using accuracy, computed over all activities performed by all individuals in the testing dataset.

*2) Qualitative evaluation:* To perform a qualitative evaluation on the ARMI task, we begin by providing an intuitive visualization of our AdHuC representation's performance, as depicted in Fig. 5(h). To emphasize our representation's impact, we compare our representation with seven approaches from previous studies, including **Harris3D** (color/depth) [17], **Cuboid** (color/depth) [18], **DSTIP** [2], **DLMC-STIP** [27] and **4D-LST** [3] representations, using the original implementation of their detectors and descriptors, as illustrated in Fig. 5.

It is observed that previous representations based on LST features are not able to extract feature affiliation information. In addition, previous methods based only upon color cues usually detect irrelevant features from dynamic background (e.g., TV) or foreground obstacles (e.g., NAO robot), while methods based on depth usually generate a large number of irrelevant features due to depth noise. Although the DSTIP method [2] avoids extracting irrelevant background features, it also fails to capture useful information from humans, especially in multi-human scenarios. As shown in Fig. 5(h), our AdHuC representation is able to identify feature affiliation, avoid extracting irrelevant features, and adapt descriptor sizes to compensate for linear perspective view changes.



(a) Color-Harris3D [17] (b) Color-Cuboid [18] (c) Depth-Harris3D [17] (d) Depth-Cuboid [18]

(e) DSTIP [2] (f) DLMC-STIP [27] (g) Color-depth 4D-LST [3] (h) **AdHuC** (color-depth)

Fig. 5. Comparison of our AdHuC representation with the state-of-the-art representations based on color-depth LST features. In Fig. 5(f), features with different colors are from different depth layers (eight layers in total). In Fig. 5(h), different feature colors denote different feature affiliations. The exemplary images are fused to clearly represent the start and end positions of the humans.

TABLE I

COMPARISON OF ACCURACY (%) AND EFFICIENCY ON THE ARMI TASK IN THE MULTI-PLAYER SIMON SAYS GAMES

| Approach | Bend | Clap | Flap | Kick | Walk | Wave | Overall | Frame rate |
|---|---|---|---|---|---|---|---|---|
| Color-Harris3D [17] | 74.2 | 73.1 | 73.4 | 71.8 | 78.2 | 76.6 | 74.5 | $\sim$ 0.267 Hz |
| Color-Cuboid [18] | 78.4 | 74.9 | 79.2 | 76.5 | 79.6 | 76.4 | 77.5 | $\sim$ 0.144 Hz |
| Depth-Harris3D [17] | 73.3 | 64.2 | 66.6 | 65.4 | 72.4 | 69.7 | 68.6 | $\sim$ 0.206 Hz |
| Depth-Cuboid [18] | 74.5 | 66.4 | 63.2 | 65.7 | 73.7 | 72.4 | 69.3 | $\sim$ 0.141 Hz |
| DSTIP [2] | 85.4 | 73.9 | 87.2 | 74.8 | **85.2** | 76.9 | 80.6 | $\sim$ 0.041 Hz |
| DLMC-STIP [27] | 75.3 | 67.2 | 69.9 | 70.8 | 75.2 | 70.5 | 71.5 | $\sim$ 0.035 Hz |
| Color+Depth 4D-LST [3] | 84.0 | 75.6 | 87.3 | 76.4 | 80.3 | 79.6 | 80.5 | $\sim$ 0.134 Hz |
| **Our AdHuC representation** | **86.7** | **78.4** | **89.4** | **80.8** | 84.9 | **82.3** | **83.8** | $\sim$ **3.267 Hz** |

*3) Quantitative evaluation:* We also conduct experiments to quantitatively evaluate our representation's performance (i.e., accuracy and efficiency). The recognition performance is presented in Table I. It is observed our AdHuC representation obtains a promising accuracy of 83.8% with a frame rate of around 3.3 Hz on the ARMI task in Simon Says games. This highlights our AdHuC approach's ability to accurately and efficiently distinguish different activities performed by multiple individuals in 3D scenes, through estimating feature affiliations.

We also compare our AdHuC representation with previous methods. Since previous approaches are not able to identify feature affiliations (and thus cannot address the ARMI task), we combine the used representation approaches with a most commonly applied baseline HOG-based human localization technique [4], [6], [7], [24], which employs a sliding window paradigm and a sparse scan using 800 local windows. The comparison is presented in Table I. It is observed that our representation outperforms the tested baselines and obtains the best overall accuracy. In addition, our algorithm significantly improves computation efficiency, as it obtains the highest frame rate. The comparison results indicate the importance of avoiding extracting irrelevant features from background in improving feature discriminative power and computational efficiency.

TABLE II

AVERAGE RECOGNITION ACCURACY ON THE MHAD DATASET

| Approach | Accuracy |
|---|---|
| Harris3D + HOG/HOF + SVM [29] | 70.07% |
| Harris3D + HOG/HOF + 3-NN [29] | 76.28% |
| Harris3D + HOG/HOF + 1-NN [29] | 77.37% |
| Depth cuboid detector and descriptor [18] | 88.72% |
| Color cuboid detector and descriptor [18] | 90.53% |
| Harris3D + HOG/HOF + MKL [29] | 91.24% |
| **Our AdHuC representation** + SVM | **97.81%** |

## B. Empirical Studies on Benchmark Datasets

To further evaluate our representation's performance, we conduct additional empirical studies on single-person activity recognition and compare our representation with approaches in previous studies, using two public color-depth benchmark datasets:

- **Berkeley MHAD** dataset [29] is a multi-model human activity dataset that contains 11 activities performed by 12 subjects in 550 instances. We make use of the rear-

TABLE III

AVERAGE RECOGNITION PRECISION ON THE ACT4$^2$ DATASET

| Approach | Precision |
|---|---|
| Harris3D + Color-HOG/HOF [30] | 64.2% |
| Depth layered multi channel STIPs + HOG/HOF [27] | 66.3% |
| Harris3D + Depth-HOG/HOF [30] | 74.5% |
| Harris3D + Comparative coding descriptor [30] | 76.2% |
| Color cuboid detector and descriptor [18] | 70.9% |
| Depth cuboid detector and descriptor [18] | 78.8% |
| Harris3D + Super feature representation [30] | 80.5% |
| **Our AdHuC representation** | **85.7%** |

view Kinect data, which were captured with a resolution of $640 \times 480$ at a frame rate of 30 Hz. Following [29], the first seven subjects are used for training and the last five for testing; accuracy is used as evaluation metric.

- **ACT4$^2$** dataset [30] is a large multi-Kinect daily action dataset containing 14 actions performed by 24 subjects in 6844 instances. The data from view four are adopted, which were captured with a resolution of $640 \times 480$ at 30 Hz. Following [30], eight human subjects are adopted for training and the remaining for testing; precision is applied as evaluation metric.

Experimental results using our AdHuC representation and comparisons with previous methods over MHAD and ACT4$^2$ datasets are presented in Tables II and III, respectively. It is observed that our representation achieves the state-of-the-art performance and significantly outperforms previous methods on single-person activity recognition from color-depth visual data. This highlights the importance of constructing human-centered representation to avoid noisy, irrelevant features.

## VI. CONCLUSION

In this paper, we introduce the novel AdHuC representation to enable intelligent robots to understand activities of multiple individuals from a sequence of 3D point clouds in practical human-centered robotics applications. To construct our representation, an affiliation region is estimated for each human through estimating depths of interest and a cascade of rejectors to localize people in 3D scenes. Then, our algorithm detects human-centered features within the affiliation region of each human, which is able to simultaneously recognize feature affiliations, avoid computing irrelevant features, and address robot movement. In addition, a new adaptive multi-channel feature descriptor is introduced to compensate for the linear perspective view variation and encode information from both color and depth channels to construct a final adaptive human-centered representation. Extensive empirical studies are performed to evaluate our AdHuC representation on a Meka humanoid robot in multiple player Simon Says gaming scenarios. Furthermore, our AdHuC representation is compared with methods in previous studies on single-person activity recognition, using MHAD and ACT4$^2$ benchmark datasets. Experimental results demonstrate that our AdHuC representation significantly improves accuracy and efficiency of human activity recognition and successfully addresses the challenging ARMI problem.

## REFERENCES


[1] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *BMVC*, 2008.

[2] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *CPVR*, 2013.

[3] H. Zhang and L. Parker, "4-dimensional local spatio-temporal features for human activity recognition," in *IROS*, 2011.

[4] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *CVPR*, 2011.

[5] S. Khamis, V. Morariu, and L. Davis, "A flow model for joint action recognition and identity maintenance," in *CVPR*, 2012.

[6] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *PAMI*, vol. 34, pp. 1549–1562, Sept. 2012.

[7] B. Ni, S. Yan, and A. A. Kassim, "Recognizing human group activities with localized causalities," in *CVPR*, 2009.

[8] I. Everts, J. C. van Gemert, and T. Gevers, "Evaluation of color STIPs for human action recognition," in *CVPR*, 2013.

[9] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzílez, "Selective spatio-temporal interest points," *CVIU*, vol. 116, pp. 396–410, Mar. 2012.

[10] O. Brdiczka, M. Langet, J. Maisonnasse, and J. Crowley, "Detecting human behavior models from multimodal observation in a smart home," *TASE*, vol. 6, pp. 588–597, Oct. 2009.

[11] M. Singh, A. Basu, and M. Mandal, "Human activity recognition based on silhouette directionality," *TCSVT*, vol. 18, pp. 1280–1292, Sept. 2008.

[12] S. Knoop, S. Vacek, and R. Dillmann, "Sensor fusion for 3D human body tracking with an articulated 3D body model," in *ICRA*, 2006.

[13] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *IEEE International Conference on Robotics and Automation*, 2012.

[14] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012.

[15] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition," in *IJCAI*, 2013.

[16] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *IJCAI*, 2013.

[17] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, pp. 107–123, Sept. 2005.

[18] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VSPETS*, 2005.

[19] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *ICME*, 2007.

[20] H. Zhang, C. M. Reardon, and L. E. Parker, "Real-time multiple human perception with color-depth cameras on a mobile robot," *TCyb*, vol. 43, no. 5, pp. 1429–1441, 2013.

[21] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *ICRA*, 2011.

[22] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, Jun. 1981.

[23] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[25] H. Grabner and H. Bischof, "On-line boosting and vision," in *CVPR*, 2006.

[26] L. Bourdev and J. Malik, "Poselets: body part detectors trained using 3D human pose annotations," in *ICCV*, 2009.

[27] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *ICCVW*, 2011.

[28] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009.

[29] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *WACV*, 2013.

[30] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *ECCVW*, 2012.