CoDe4D: Color-Depth Local Spatio-Temporal Features for Human Activity Recognition From RGB-D Videos

Hao Zhang, *Member, IEEE*, and Lynne E. Parker, *Fellow, IEEE*

***Abstract*— Human activity recognition has a variety of important real-world applications, such as video analysis, surveillance, and human-robot interaction. As a promising video representation method, local spatio-temporal (LST) features have received increasing attention from computer vision, machine learning, and robotics communities. However, approaches based on traditional LST features only use color information, which face several challenges, such as illumination changes and dynamic backgrounds. The recent availability of commercial color-depth cameras makes it much cheaper, faster, and easier to acquire depth information, which provides a potential to implement more discriminative and robust LST features. In this paper, we introduce the new 4-D color-depth (CoDe4D) LST feature that incorporates both intensity and depth information acquired from RGB-D cameras. Our feature detector constructs a saliency map through applying independent filters in $xyzt$ dimension to represent texture, shape and pose variations, and selects its local maxima as interest points. Our multichannel orientation histogram descriptor applies a 4-D support region, which is adaptive to linear perspective view changes, on each interest point. Then, image gradients of color-depth patches within the support region are computed and quantized using a spherical coordinate-based method to form a final feature vector. We build a complete activity recognition system by combining our features with bag-of-features representations and support vector machines. To evaluate the performance of our CoDe4D LST features and the complete system, we conduct experiments using four benchmark color-depth human activity data sets, including UTK Action3-D, Berkeley MHAD, ACT4^2, and MSR daily activity 3-D data sets. Experimental results demonstrate the promising representative power of our CoDe4D features, which obtain the state-of-the-art performance on activity recognition from RGB-D visual data.**

***Index Terms*— 3-D machine vision, activity recognition, color-depth feature, human representation, RGB-D sensor.**

I. Introduction

In RECENT years, human activity recognition in unstructured environments has drawn increasing attention from researchers in the field of computer vision, machine learning, and robotics [1]–[3]. This interest is mainly driven by a large number of important real-world applications. For example, in surveillance and security applications [4], automatic recognition of human activities is important to identify potential criminal and dangerous behaviors; in human-centered robotics [5], an activity recognition system allows intelligent robotic systems to better serve humans in human social environments. However, vision-based human activity recognition in real-world scenarios is a challenging task due to variations of human poses, diversity of human appearance, and traditional computer vision difficulties, such as camera movement, dynamic background, illumination variation, occlusion, and so on.

Among different approaches for representing human activities [6]–[9], local spatio-temporal (LST) features have recently become a most popular representation, which are inspired by a human visual attention mechanism that allows a human to use salient body appearances and movements to rapidly recognize activities in complex, cluttered scenes [10]. LST features are designed to capture variations of characteristic textures, shapes, and poses in visual data and thereby to provide a descriptive representation of human activities in a video. These features are typically defined as spatio-temporal pixels, referred to as interest points, which maximize a user-defined saliency function. LST features are often described using local appearance and motion information in the neighborhood of each selected interest point. Since LST features are relatively invariant to image rotation, scaling, and translation, partially invariant to illumination changes, and robust to partial occlusion [11], [12], they are widely used to encode human activities in color videos [5], [13]–[18]. In addition, because LST features are directly extracted from raw visual data, they avoid potential failures of preprocessing steps, such as human detection and tracking.

Although LST features have shown promising performance on human activity recognition from color videos, because of the limitation of the sensing device (e.g., color cameras), most of previous LST features do not make use of one important piece of information that is now available—depth. Because humans act in 3-D space, depth can be utilized along with color cues to implement more distinctive, robust salient features. The depth sensor has several advantages over color cameras. First, depth sensors provide 3-D structure information of the scene, which significantly alleviates the limitation of traditional vision systems that only acquire information

Manuscript received January 25, 2014; revised May 8, 2014 and October 15, 2014; accepted November 15, 2014. Date of publication November 26, 2014; date of current version March 3, 2016. This paper was recommended by Associate Editor X. Wang.

H. Zhang is with the Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401 USA (e-mail: hzhang@mines.edu).

L. E. Parker is with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996 USA (e-mail: leparker@utk.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2014.2376139

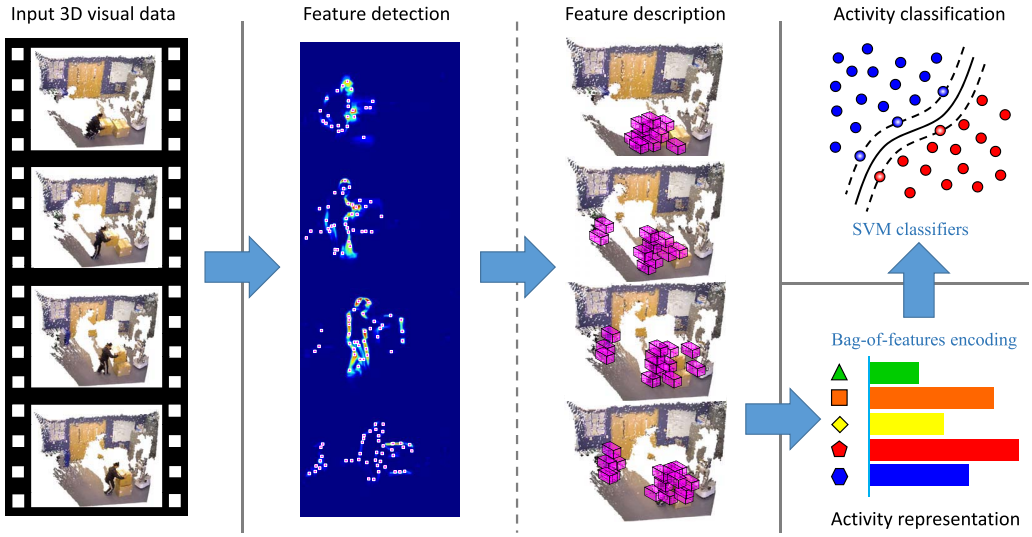


Fig. 1. Major components to recognize human activities using our CoDe4D LST features. Given a sequence of 3-D (xyz) frames (e.g., 3-D point clouds or color-depth images), our feature detection algorithm constructs a sequence of saliency maps, which characterize texture, shape, and pose variations using both color and depth information, and extracts STIPs from the saliency map sequence. Then, our multichannel feature description algorithm centers an adaptive support region at each interest point to incorporate information in its neighborhood and encodes intensity and depth image gradients within the support region to form a feature vector. To recognize human activities using the proposed CoDe4D LST features, we construct a complete system using the BoFs representation and the SVM classifier.

in 2-D space. Second, depth sensors are generally not sensitive to illumination changes and can work in darkness, which allows for obtaining observations at night. Because of the emergence of affordable color-depth cameras, such as Microsoft Kinect, Asus Xtion Pro Live, and PrimeSense color-depth sensors, it is now much faster, easier, and cheaper to construct a 3-D vision system that is able to perceive humans in 3-D space.

In this paper, we introduce the 4-D color-depth (CoDe4D) LST features that are extracted in $xyzt$ space (i.e., 3-D spatial and 1-D temporal) and incorporate both color and depth information contained in a sequence of RGB-D frames or 3-D point clouds. The objective of developing such CoDe4D features is to provide a compact, discriminative representation of human activities when depth information is available along with color information, in order to improve activity recognition performance in realistic, complex scenes.

An overview of our CoDe4D LST feature extraction method along with how the features are employed to construct a human activity recognition system is graphically summarized in Fig. 1. Given a sequence of 3-D point clouds or color-depth frames, we construct a spatio-temporal saliency map that incorporates both color and depth information, and define interest points as the local maxima of the saliency map. Then, we place a 4-D hyper-cuboid as the feature's support region at each detected interest point in 4-D (i.e., 3-D space and 1-D time), which adapts its size to the interest point's depth value; then we use the Multichannel Orientation Histogram (MCOH) descriptor to incorporate color and depth cues to form a final feature vector. To perform human activity recognition using our CoDe4D features, we apply the standard bag-of-features (BoFs) model, which quantizes our CoDe4D LST features into discrete visual words and represents each input sequence as a frequency histogram of the words. Then, a nonlinear support vector machine (SVM) with a χ^2-kernel is applied to perform multiclass activity classification.

The contributions of this paper are summarized as follows.

1) We propose a new CoDe4D multichannel feature detector based on a salience map, which considers both color and depth cues to extract LST interest points in $xyzt$ space.
2) We implement a new feature descriptor, called MCOH, which is able to encode both color and depth cues and adapt the support region size to visual linear perspective variations.
3) We empirically validate that our CoDe4D LST features extracted using the multichannel detector and MCOH descriptor are highly discriminative to represent human activities, which results in state-of-the-art activity recognition performance.

This paper is based on and significantly extends our previous work discussed in [5]. In particular, the new MCOH descriptor is implemented, which is more effective and efficient than the naive flat feature descriptor proposed in [5]. In addition, the CoDe4D features are comprehensively evaluated using four benchmark data sets under the supervised learning framework in this paper, which is different from [5] that evaluated feature performance only based on one data set under the unsupervised learning framework. Finally, additional explanations and illustrations of our CoDe4D feature detector and MCOH descriptor are provided in this paper.[1]

The remainder of this paper is organized as follows. In Section II, we provide a comprehensive overview of previous features to represent human activities in 3-D space. Our multichannel feature detector to detect the CoDe4D

[1]Additional information is available at http://dilab.eecs.utk.edu/CoDe4D.

LST features is proposed in Section III. Then, the MCOH descriptor applied to quantize our CoDe4D LST features is introduced in Section IV. Section V briefly describes our activity recognition approach based on BoF models and SVMs. Evaluation results of our CoDe4D LST features for activity recognition tasks are presented in Section VI. Finally, the conclusion is drawn in Section VII.

II. Related Work

A large number of features have been proposed to represent and recognize activities from visual data [1]–[3]. We review different categories of feature extraction methods with a focus on approaches of working with 3-D visual data and LST features.

A. Activity Representation in 3-D Space

Although most previous features for activity representation are based on 2-D videos [16], [17], [19], [20], several methods using 3-D visual data were proposed in the past few years, which can be generally classified into four groups. A naive human activity representation is based on the 3-D centroid trajectory, in which a human subject is represented as a point that indicates the 3-D location of the human subject in the visual data [21], [22]. In general, features based on the centroid trajectory are only suitable for representing a human who occupies a small region in an image. Another representation of human activities using 3-D visual data is based on human shape information, including a history of 3-D human silhouette [23]–[25]. A third category of representation to recognize human activities is based on 3-D human models, such as a 3-D human skeleton model [26] and a 3-D articulated body-part model [27]–[29]. The robustness of the features based on 3-D human shapes and body models relies heavily on the performance of foreground human segmentation and body part tracking, which are hard-to-solve problems due to camera motions, dynamic background, and occlusions [30]. Different from the discussed three categories of features that are extracted globally from 3-D visual data, the last category of features do not require global information (e.g., human locations) to compute.

B. LST Feature Detection

LST features, which represent global human activities with local texture, shape, and pose changes, have recently become a most popular activity representation due to their promising performance on human activity classification. The first LST feature detector, referred to as spatio-temporal interest point (STIP) detector, was introduced in [14], which is based on generalized Harris corner detectors with a set of multiscale spatio-temporal Gaussian derivative filters. Dollár *et al.* [13] detected LST features, often referred to as cuboid features, from color videos through applying separable filters in spatial and temporal dimensions (i.e., Gaussian filters along spatial dimension and Gabor filters along temporal dimension) and selecting interest points with maximum responses in the motion saliency map. Other LST features were also developed based on an extended Hessian saliency measure [31], a salient region detector or global information [32].

These approaches extract LST features only based on color information and ignore the important depth information that is available in color-depth videos. Recently, several features were introduced to extract LST features from depth images. Cheng *et al.* [33] applied the STIP detector directly on depth images obtained from color-depth sensors. Ni *et al.* [34] introduced the depth-layered multichannel STIPs (DLMC-STIPs) by applying the standard STIP detector on multiple depth layers. Xia and Aggarwal [9] proposed the depth-STIPs through applying the cuboid detector on depth images. Although these feature detection methods can extract visual cues from depth images, they do not make use of color or intensity information and therefore ignore important texture information. Different from previous feature detectors that are based on either color [13], [14] or depth [9], [33] cues, we introduce a multichannel LST feature detector that is capable of incorporating both color and depth information during the detection process and extract LST features from color-depth visual data.

C. LST Feature Description

After an interest point is detected, a descriptor is required to encode the information in the neighborhood of the interest point to construct a final feature vector. Nearly, all LST feature descriptors used to represent human activities in color videos are based on image gradients. Dollár *et al.* [13] concatenated gradients of intensity images into a feature vector. Scovanner *et al.* [35] implemented the scale-invariant feature transform (SIFT) 3-D descriptor, an extension of the well-known SIFT [11] descriptor to color videos, to describe gradients in space–time dimensions using spherical coordinate-based quantization methods. Kläser *et al.* [36] implemented the histogram of oriented gradient (HOG) 3-D descriptor, an extension of the well-known HOGs descriptor [37], to describe spatio-temporal gradients computed from color image sequences using regular polyhedron-based quantization approaches. Laptev *et al.* [38] implemented the HOG/histogram of optimal flow (HOF) descriptor to characterize local shapes and motions for activity recognition from color videos. Another popular type of feature descriptors investigates trajectories of interest points based on optical flow [17], [18]. The low-level local features were also aggregated to construct more complicated middle-level human activity representations, such as the motionlet [39].

With the emergence of depth sensors (e.g., Kinect), several LST descriptors were developed to quantize local cues contained in depth image sequences [40]. A most commonly used methodology to describe visual features from depth frames is to extend the existing color/intensity descriptors. For example, Ni *et al.* [34] directly applied the HOG/HOF descriptor on their DLMC-STIPs; Ofli *et al.* [41] also employed HOG/HOF descriptors to quantize features detected from depth frames for activity recognition. To represent depth information for human action analysis, Cheng *et al.* [33] introduced the comparative coding descriptor, which describes

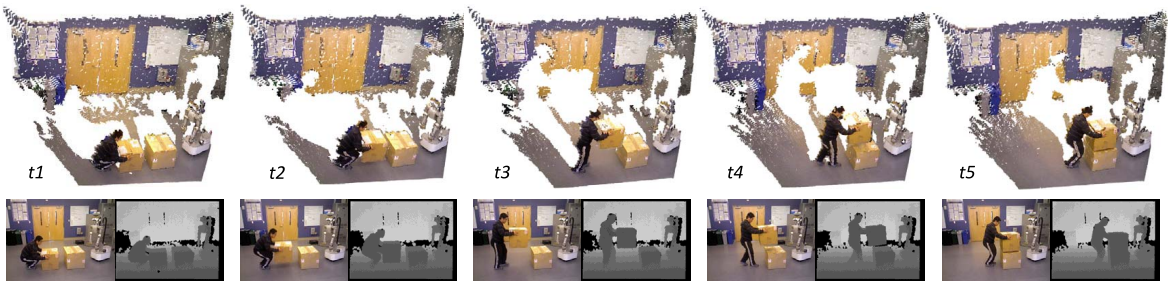


Fig. 2. Exemplary input sequence of 3-D frames (i.e., point clouds and their respective color-depth frames). Noise reduction is applied on the visual data, which aligns color and depth pixels, reduces auto-balancing fluctuation and removes depth noise. In this example, a human subject is performing a box-lifting activity in a human-robot collaboration application.

the structure of the depth cuboid using sequential codes. Xia and Aggarwal [9] developed the depth cuboid similarity feature descriptor that uses self-similarity to encode spatio-temporal shapes of depth cuboids. Oreifej and Liu [42] applied surface normals to describe local spatial cues. The low-level features were also applied to construct more complicated representations, such as the super normal vector [43].

Previous feature description algorithms are based on either color or depth cues. Different from these descriptors, we aim at developing a multichannel descriptor that can simultaneously encode both color and depth cues and adapt to linear perspective view variations. To this end, we significantly improve our previous descriptor introduced in [5] by adapting the support region size and designing spherical coordinate-based methods to quantize visual cues that are extracted from both color and depth channels.

III. CoDe4D Feature Detection

Color-depth data obtained from RGB-D cameras generally contains massive amounts of information in the form of spatio-temporal color and depth variations. Most of the information, such as pixels representing floors and background clusters, is not directly relevant to human subjects or informative enough to represent human activities. Accordingly, it is highly desirable to extract compact, discriminative features from color-depth visual data to effectively encode human activities. Our CoDe4D LST feature is introduced to address this important problem. The feature is defined in 4-D space in the sense that it characterizes local pose, shape, and texture variations in 3-D spatial dimension (i.e., xyz) and 1-D temporal dimension (i.e., t).

A. Noise Reduction

Color-depth visual data obtained from the RGB-D camera usually contains a considerable amount of noise. Accordingly, noise reduction is an important process before extracting LST features. We identify three major noise sources.

1) *Color-Depth Misalignment:* RGB-D cameras acquire color and depth information independently; as a consequence, the obtained color and depth images can be misaligned. To reduce this misalignment noise, a color-depth camera should be calibrated by adjusting its intrinsic parameters, such as focal distances, distortion coefficients, and image centers, in order to accurately map between depth pixels and color pixels. For example, as shown in Fig. 2, the depth pixels are mapped to their respective color pixels.
2) *Improper Auto White Balance:* The color sensor of RGB-D cameras uses an auto white balance mechanism, which usually causes a significant fluctuation of the RGB value of a pixel under minor variations in the light. To handle this noise source, histogram equalization is applied over RGB images to reduce the white balance fluctuation.
3) *Depth Sensing Defect:* The depth sensor of RGB-D cameras captures depth by projecting discrete infrared (IR) patterns on the scene and measuring their displacement. Due to the limitation of this depth sensing technology, acquired depth data often contains a large number of pixels with missing values, which can result from occlusions of the depth camera's point of view or the absorption of the IR light by objects. To handle this type of noise, erosion and dilation [44] are performed to remove noisy pixels and small structures in depth images; then, hole filling using morphological reconstruction [44] is applied on black regions to estimate depth for pixels with missing depth values.

The resulting color-depth visual data serves as the input to our multichannel LST feature extraction algorithm to compute the CoDe4D LST features in $xyzt$ space.

B. Spatio-Temporal Filtering

We denote the color-depth visual data (e.g., 3-D point cloud sequences or color-depth videos) as a sequence of 3-D frames $\{I_1, \ldots, I_T\}$. The 3-D frame at time point t is denoted by $I_t = (x, y, z, i, t)$, $\forall t \in [1, T]$, where x and y represent pixel locations in the image; z is the pixel's depth value in the range of 0–255 that is typically mapped from physical range of 0–8 m obtained by depth cameras; and i is the intensity value computed from its respective RGB values. It is noteworthy that depth values can be considered as a function $\mathbb{R}^3 \rightarrow \mathbb{R} : z = z(x, y, t)$, which constitutes a hyper-surface in 4-D space represented as $S(x, y, t, z(x, y, t)) = 0$.

The first step to detect LST interest points is to incorporate space-time information. To achieve this objective,

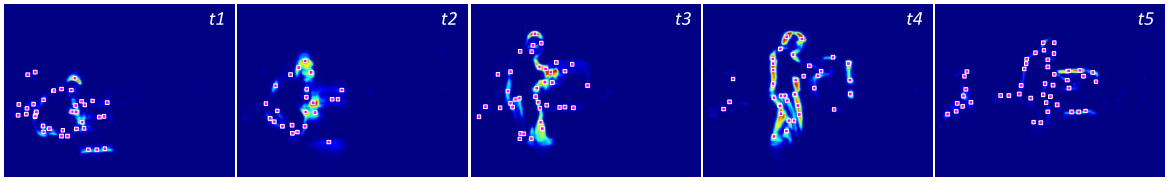


Fig. 3. Spatio-temporal saliency map that combines both color and depth information to characterize space-time variations of poses, textures, and shapes. Warmer colors in the saliency map represent stronger variations. Our STIPs are defined as the local maxima, which have the strongest local variations, of the saliency map in xyt space, as depicted by the magenta boxes with white edges.

we propose a separable filtering algorithm that employs independent 3-D spatial and 1-D temporal filters on each intensity-depth pixel to consider spatio-temporal variations in $xyzt$ space. To incorporate spatial variations, a pass-through filter and a Gaussian filter are applied to spatially smooth intensity and depth values of each 3-D frame along xyz dimensions

$$i_s(x, y, t) = (i(x, y, t) \circ f(z|\delta)) * p(x, y|\sigma) \quad (1)$$

$$z_s(x, y, t) = (z(x, y, t) \circ f(z|\delta)) * p(x, y|\sigma) \quad (2)$$

where $*$ denotes convolution, and $\circ$ represents Hadamard product (entry-wise matrix multiplication). $f(z|\delta)$ is the pass-through filter parameterized by δ, which controls the spatial scale along the depth dimension and is applied to prune pixels falling outside of the depth range

$$f(z|\delta) = \mathbb{1}(|z(x, y, t) - z)| \leqslant \delta). \quad (3)$$

The function $p(x, y|\sigma)$ is a 2-D Gaussian filter applied along x and y spatial dimensions. The parameter σ of the Gaussian filter controls its spatial scale

$$p(x, y|\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{\|x^2+y^2\|}{2\sigma^2}}. \quad (4)$$

To combine variations of intensity-depth pixel values across frames, a Gabor filter is applied along the time dimension over the spatially filtered 3-D frames

$$i_{\text{st}}(x, y, t) = i_s(x, y, t) * g(t|\tau, \omega) \quad (5)$$

$$d_{\text{st}}(x, y, t) = d_s(x, y, t) * g(t|\tau, \omega) \quad (6)$$

where $g(t|\tau, \omega)$ is a complex-valued Gabor filter given by

$$g(t|\tau, \omega) = \frac{1}{\sqrt{2\pi}\tau} \cdot e^{-\frac{t^2}{2\tau^2}} \cdot e^{i(2\pi\omega t)} \quad (7)$$

where τ controls the temporal scale of our feature detector. Throughout this paper, we assign $\omega = 0.6/\tau$, which empirically shows good human activity representation and classification performance.

C. Interest Point Detection

In order to identify space-time interest points, we construct a spatio-temporal saliency map from the responses of intensity and depth filters, as

$$r(x, y, t) = (1 - \alpha) \cdot \|i_{\text{st}}(x, y, t)\|^2 + \alpha \cdot \|d_{\text{st}}(x, y, t)\|^2 \quad (8)$$

where α is a mixture weight to balance between intensity and depth information. The spatio-temporal saliency map generally represents variations of textures, shapes, and poses, because any region undergoing such variations induces responses. For example, the saliency map of the box-lifting activity in Fig. 2 is shown in Fig. 3, where warmer colors denote stronger responses, indicating that the frame has larger texture, shape, and pose variations. It is noteworthy that our saliency map is defined in 3-D space, which encodes variations of pixel values in xyt space.

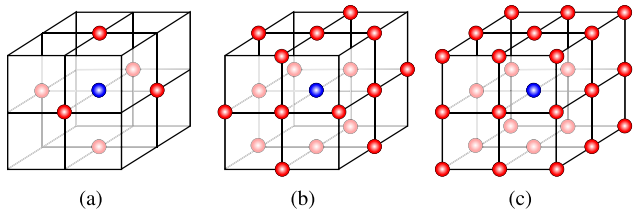


Fig. 4. Pixel connectivity defined to compute the local maxima of the 3-D saliency map. The blue dot denotes the query pixel; the red dots represent its connected neighbors. (a) Six connected. (b) 18 connected. (c) 26 connected.

Given the saliency map, our STIPs are defined as the local maxima of the map; that is, the pixels having the most significant variations. We employ an approach based on connected neighbors to compute local maxima of the 3-D saliency map, using six neighbors [pixels that touch one of the faces of the query pixel, as shown in Fig. 4(a)], 18 neighbors [pixels that touch one of the faces or edges, as shown in Fig. 4(b)], or 26 neighbors [pixels touching one of the faces, edges, or corners, as shown in Fig. 4(c)]. As an example, the STIPs detected from the saliency map of the box-lifting activity (as shown in Fig. 2) are shown in Fig. 3, which are computed based on 18-connected neighbor pixels. Because the introduced feature detection algorithm is able to incorporate both color and depth information to select LST interest points in $xyzt$ space, our detector is referred to as the CoDe4D LST feature detector.

D. Computational Complexity

The running time of our CoDe4D feature detection method is $O(\text{MN}\sigma^2) + O(\text{MN}\tau)$, where the first term $O(\text{MN}\sigma^2)$ comes from applying 2-D Gaussian filters and second term is from using the Gabor filter. Typically, $\sigma^2 \gg \tau$ holds, resulting in an average runtime of $O(\text{MN}\sigma^2)$.

IV. CoDe4D Feature Description

After interest points are detected in space-time dimensions, which represent locations of our CoDe4D features, a feature

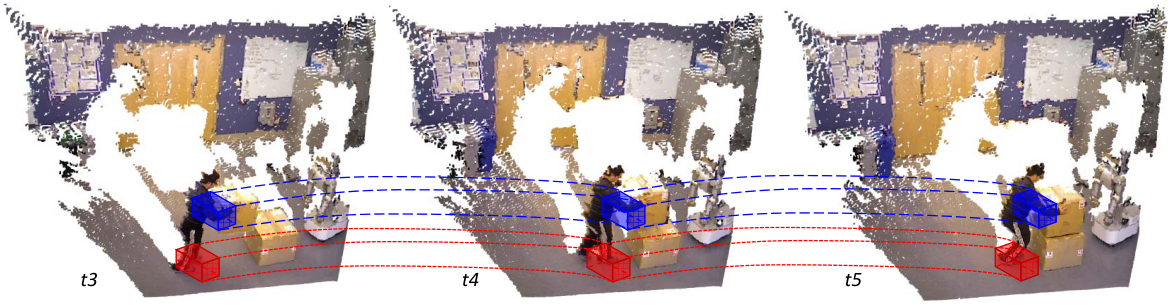


Fig. 5. Spatio-temporal support regions in 4-D ($xyzt$) space, which are used in our feature description algorithm to incorporate color-depth information in the neighborhood of detected interest points. Given a STIP, which is detected within a specific time span, its 4-D support region is constructed by centering a 3-D (xyz) cuboid at the point of each frame within the time span (e.g., the blue and red cubes in each 3-D frame) and connecting these 3-D cuboids across multiple frames along 1-D time dimension (as illustrated by the dashed lines across multiple 3-D frames).

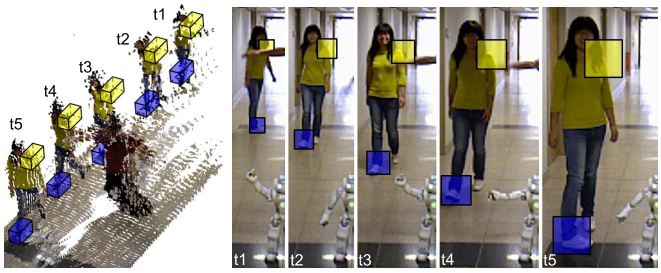


Fig. 6. Linear perspective view changes. Support regions that have the same size in 3-D (xyz) physical space have different projected sizes when they are mapped onto 2-D (xy) images, due to linear perspective view changes, as shown by the smaller support regions on 2-D images when the human subject performs activities further away from the camera.

descriptor is required to incorporate the information contained in the neighborhood of each detected interest point in order to form a final feature vector. The neighborhood of an interest point is typically encoded by a support region that is centered at the point. In this paper, we define our support region $\boldsymbol{S}$ as a hyper-cuboid in $xyzt$ dimensions, which is parameterized by an octuple, i.e., $\boldsymbol{S} = (x, y, z, t, s_x, s_y, s_z, s_t)$, where (x, y, t) is the location of the STIP extracted from the saliency map in xyt space, z is the depth value of the interest point, i.e., $z = z(x, y, t)$, and (s_x, s_y, s_z, s_t) represents the support region's size along 3-D spatial and 1-D temporal dimensions. Examples of the 4-D support regions in $xyzt$ space are shown in Fig. 5.

A. Adaptive Support Region

Based only on color cues, adapting the support region's size is generally a hard-to-solve problem due to the difficulties of estimating depth values from color cues. Taking advantage of the color-depth sensing technology, we can use the available depth information provided by the depth sensor to estimate 3-D geometry structures of a scene, and thereby adapt the size of a support region to linear perspective view variations, i.e., an object closer to the camera seems to have a larger size. This phenomenon is shown in Fig. 6. When the human is walking toward the color-depth camera, the size of the support regions should remain the same in 3-D (xyz) physical space. This is because these support regions are used to incorporate information contained in local regions, such as left shoulder and right foot of the human in Fig. 6, whose size is generally not changed. However, when the support regions are mapped onto 2-D images, their size is changed due to linear perspective view changes. In order to address this important but not well studied issue, our adaptive support region is introduced.

Since LST interest points are usually detected on boundaries (e.g., corners and edges), a number of detected points can fall out of a human blob (i.e., a region of the 3-D frame that only contains pixels from a human subject), even though they are generated by humans and represent human pose, texture and shape variations. To handle this issue, we introduce a method to estimate the more accurate depth of an interest point, with the objective to adapt support region sizes to linear perspective view changes. Given scales of the space-time filters $(\sigma, \sigma, \delta, \tau)$ that are applied to detect interest points, for each interest point located at (x, y, z, t) (where z can be inaccurate if the interest point falls out of human blobs), we estimate a more accurate depth for the point using the following two steps.

1) Construct the spatial-temporal detection cuboid in $xyzt$ space $\boldsymbol{C} = (x, y, z_t, t, 2\sigma, 2\sigma, 2\delta, 2\tau)$, which is centered at (x, y, z_t, t), where $z_t = z(x, y, t)$ is the depth value of the pixel (x, y) at time t.
2) Estimate a new depth value for the point (x, y, z_t, t) by calculating the minimum depth of the points within the spatio-temporal detection cuboid $\boldsymbol{C}$, which is mathematically defined as

$$z(\boldsymbol{C}) = \min_{\substack{z\in[z_t-\delta, z_t+\delta] \\ \forall i\in[x-\sigma, x+\sigma] \\ \forall j\in[y-\sigma, y+\sigma] \\ \forall k\in[t-\tau, t+\tau]}} z(i, j, k). \tag{9}$$

Then, the estimated depth value $z(\boldsymbol{C})$ is used as the depth of the interest point. Our depth estimation approach is based on the plausible assumption of foreground humans, which is a typical situation for most color-depth camera applications in indoor environments, such as gaming [45] and human-robot social interaction [46]. The plausibility of the assumption can also be observed from benchmark color-depth human activity data sets, including UTK Action3-D [5], Berkeley MHAD [41], ACT4^2 [33], and MSR daily activity 3-D [47] data sets, in which human subjects always stay in the foreground.

After estimating $z(\boldsymbol{C})$, the support region is placed at $z(\boldsymbol{C})$ in the depth dimension, i.e., $\boldsymbol{S} = (x, y, z(\boldsymbol{C}), t, s_x, s_y, s_z, s_t)$. Then, we can adapt the size of the support region to compensate for linear perspective view changes along xy dimensions, as

$$s_x = s_y = \frac{\sigma_0 \sigma}{z(\boldsymbol{C})} \tag{10}$$

where σ_0 characterizes the support region's relative spatial size along xy dimensions. Since the depth dimension is not affected by the linear perspective view variation, we define $s_z = \delta_0 \delta$, where δ_0 encodes the relative spatial size in the z dimension. Similarly, the support region's temporal size is not affected by spatial linear perspective view variations, and thus we define $\tau_s = \tau_0 \tau$, where τ_0 characterizes the relative temporal size.

B. Multichannel Orientation Histogram

We introduce a multichannel (color and depth) descriptor, based on image gradient orientations, to quantize visual cues within a support region in $xyzt$ space. Because a visual cue's orientation is independent of its magnitude that is affected by image noise and illumination changes, orientation quantization has proved to be a robust methodology for feature description [11], [12], [35], [37].

Given the support region $\boldsymbol{S}$ in $xyzt$ space that contains a set of pixels with intensity-depth values, we first decompose $\boldsymbol{S}$ into sequences of intensity and depth image patches in xyt space, i.e., $i_p(x, y, t)$ and $z_p(x, y, t)$. Then, we compute spatio-temporal gradients of the intensity patch sequence along x, y, and t dimensions as

$$\nabla i_p = \left(\frac{\partial i_p}{\partial x}, \frac{\partial i_p}{\partial y}, \frac{\partial i_p}{\partial t} \right) \tag{11}$$

where the gradient along each dimension is computed using the finite difference approximation

$$\begin{aligned} \frac{\partial i_p(x, y, t)}{\partial x} &= i_p(x+1, y, t) - i_p(x-1, y, t) \\ \frac{\partial i_p(x, y, t)}{\partial y} &= i_p(x, y+1, t) - i_p(x, y-1, t) \\ \frac{\partial i_p(x, y, t)}{\partial t} &= i_p(x, y, t+1) - i_p(x, y, t-1). \end{aligned} \tag{12}$$

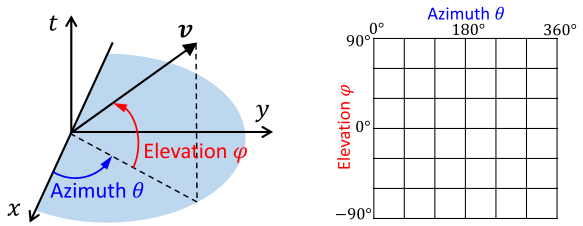
Spatio-temporal gradients of the depth image patch sequence, i.e., ∇z_p, can be computed in the same way.

We quantize the gradients of image patch sequences in the support region using a spherical coordinate-based approach. For each spatio-temporal intensity image gradient vector, we compute its azimuth $\theta(\nabla i_p)$ and elevation $\varphi(\nabla i_p)$ angles to characterize its 3-D orientations in xyt space, as

$$\theta(\nabla i_p) = \arctan \frac{\partial i_p}{\partial y} \Big/ \frac{\partial i_p}{\partial x} \tag{13}$$

$$\varphi(\nabla i_p) = \arctan \frac{\partial i_p}{\partial t} \Big/ \sqrt{\frac{\partial^2 i_p}{\partial y} + \frac{\partial^2 i_p}{\partial x}}. \tag{14}$$

An intuitive explanation of azimuth and elevation computation is shown in Fig. 7(a). Then, the azimuth θ and elevation φ angles of each image gradient are discretized using 2-D bins, as graphically explained in Fig. 7(b). Finally, a 1-D histogram is formed through concatenating all entries of the 2-D bins. A histogram of 3-D gradient orientations of depth patch sequences can be computed using the same procedure.



(a) (b)

Fig. 7. 3-D feature description based on spherical coordinates. Each 3-D gradient is decomposed into two independent azimuth and elevation angles, as shown in (a). Then, the angles are quantized via binning, as depicted by the example in (b), which subdivides azimuth and elevation angles into six bins each, leading to a histogram of 36 bins. (a) Orientation computation. (b) Orientation quantization.

In order to construct a final feature vector that contains both intensity and depth information, we implement the MCOH descriptor, based on the histograms of the intensity and depth image patch gradient orientations. To deal with adaptive support region size, which can lead to a different number of elements in the histograms, we apply normalization on the histograms. In particular, given the histograms of intensity and depth gradient orientations, $\boldsymbol{h}_i$ and $\boldsymbol{h}_z$, the final feature vector $\boldsymbol{h}$ is constructed by

$$\boldsymbol{h} = \left(\frac{\boldsymbol{h}_i}{2N_i}, \frac{\boldsymbol{h}_z}{2N_z} \right) \tag{15}$$

where N_i and N_z are the total number of gradient orientations in $\boldsymbol{h}_i$ and $\boldsymbol{h}_z$, respectively.

C. Computational Complexity

The worst case time complexity of our CoDe4D descriptor is $O(\delta^2 \tau(N_c))$, where N_c is the number of features from each color-depth frame. The support region of a feature contains at most $\delta^2 \tau$ points since color-depth cameras can only capture a surface of points in xyz space.

V. Human Activity Recognition

A. Representation

We apply the standard BoFs representation to encode visual data as well as human activities, which is the most widely used representation based on LST features [12], [36]. An overview of our BoF encoding method is graphically shown in Fig. 8.

The BoF representation requires a visual vocabulary. To this end, we construct our vocabulary using clustering, which is shown to be robust against scale changes and camera motions [48]. We employ the standard k-means algorithm to cluster a subset of randomly selected CoDe4D features. Each cluster is indexed by a visual word. Then, each feature is assigned to its nearest visual word using Euclidean distance.

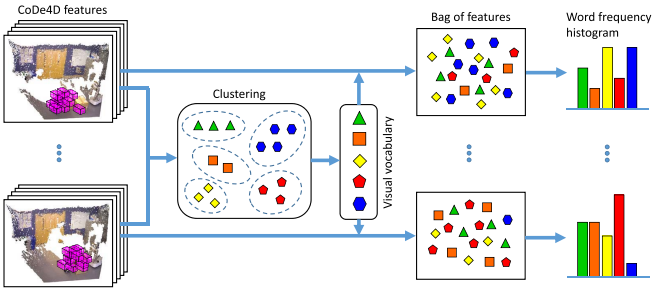


Fig. 8. BoFs encoding for video representation. CoDe4D LST features extracted from color-depth visual data are clustered to construct a visual vocabulary, with each cluster representing a group of similar features. Then, each color-depth data instance is represented using the BoFs model, which is eventually encoded as a histogram of visual word frequency.

During clustering, random feature selection is used to reduce computational complexity; we also execute the k-means algorithm multiple times using different initializations to obtain a vocabulary that has the lowest error (i.e., within-cluster sum of squares).

Vocabulary construction is a most important component in the BoF representation, since it can significantly reduce feature dimensions: each feature vector is encoded by a single visual word. Then, each instance of visual data (e.g., a sequence of 3-D point clouds or color-depth frames) can be represented as a histogram of visual word occurrences.

B. Classification

We apply SVMs as a benchmark classifier to perform activity recognition. In order to deal with the discrete BoF representation, which serves as the input to SVMs, we use the χ^2-kernel [49]. Given two histograms $\boldsymbol{h}_a = \{h_{ak}\}$ and $\boldsymbol{h}_b = \{h_{bk}\}$, the kernel is computed by

$$K(\boldsymbol{h}_a, \boldsymbol{h}_b) = \exp\left(-\frac{1}{A} D(\boldsymbol{h}_a, \boldsymbol{h}_b)\right) \tag{16}$$

where $D(\cdot)$ is the χ^2-distance defined as

$$D(\boldsymbol{h}_a, \boldsymbol{h}_b) = \frac{1}{2} \sum_k \frac{(h_{ak} - h_{jk})^2}{h_{ak} + h_{jk}} \tag{17}$$

and A is a constant denoting the average χ^2-distance between all pairs of N training instances

$$A = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} D(\boldsymbol{h}_i, \boldsymbol{h}_j). \tag{18}$$

For multiclass activity classification, the standard one-against-one methodology is employed.

VI. Experiments

In this section, we detail the empirical study performed to evaluate our CoDe4D LST features' performance on recognizing human activities from color-depth visual data.

A. Implementation

In our CoDe4D LST feature detector, we set the spatial scale along xy dimensions to $\sigma = 5$ pixels and $\delta = 0.3$ meters in the z dimension, and assign the temporal scale to $\tau = 3$ frames. We apply the color-depth mixture parameter $\alpha = 0.75$. We use 18-connected neighbors to select local maxima of the saliency map in xyt space. These parameter values can result in satisfactory activity recognition performance in general situations. Further explanations regarding the parameter selection process will be discussed in Section VI-D.

When implementing our adaptive feature support region, we set the relative spatial sizes to $\sigma_0 = \delta_0 = 5$ and the relative temporal size to $\tau_0 = 4$. In our MCOH descriptor, we divide elevation angle φ into 6 bins and azimuth angle θ into 12 cells, resulting in a final feature vector containing 72 elements.

In the BoF encoding, we randomly select 100 000 CoDe4D LST features extracted from the training set of a given data set to construct a vocabulary containing 2000 visual words, which empirically shows promising activity recognition performance (as will be discussed in Section VI-D). The vocabulary construction process is repeated eight times using different initializations; the result with the minimum clustering error is selected as our final vocabulary. A total number of 500 CoDe4D LST features that have the largest values in the saliency map are used to construct the histogram of word occurrences from each data instance. The histogram is used as input to SVMs.

B. Data Sets

We use four benchmark color-depth human activity data sets to evaluate our feature's performance on activity recognition: the UTK Action3-D data set to follow our previous work [5], and the Berkeley MHAD, ACT4[2] and MSR daily action 3-D data sets, which are the modern large-scale color-depth activity data sets. Details of the data sets are listed as follows.

1) *UTK Action3-D:* Data set[2] [5], is an earliest RGB-D human activity data set that is publicly available. This data set contains six human activities, including sequential activities, repetitive activities, and activities with small movements. Each activity class contains 33 instances. Each instance consists of a color video and a calibrated depth video. This RGB-D activity data set was collected using a Kinect sensor that is installed on a Pioneer 3-DX mobile robot in human social environments, such as office and home. The UTK Action3-D data set contains many challenges, including illumination changes, dynamic background, and variations in human appearances and motions. Exemplary frames of the data set are shown in Fig. 9(a).
2) *Berkeley MHAD:* Data set[3] [41] is a multimodal human activity data set that contains 11 activities performed by 7 male and 5 female subjects. Each activity was repeated five times, yielding ~550 activity video sequences.

[2]The UTK Action3-D data set is available at http://dilab.eecs.utk.edu/CoDe4D.

[3]The Berkeley MHAD data set is available at http://tele-immersion.citris-uc.org/berkeley_mhad.

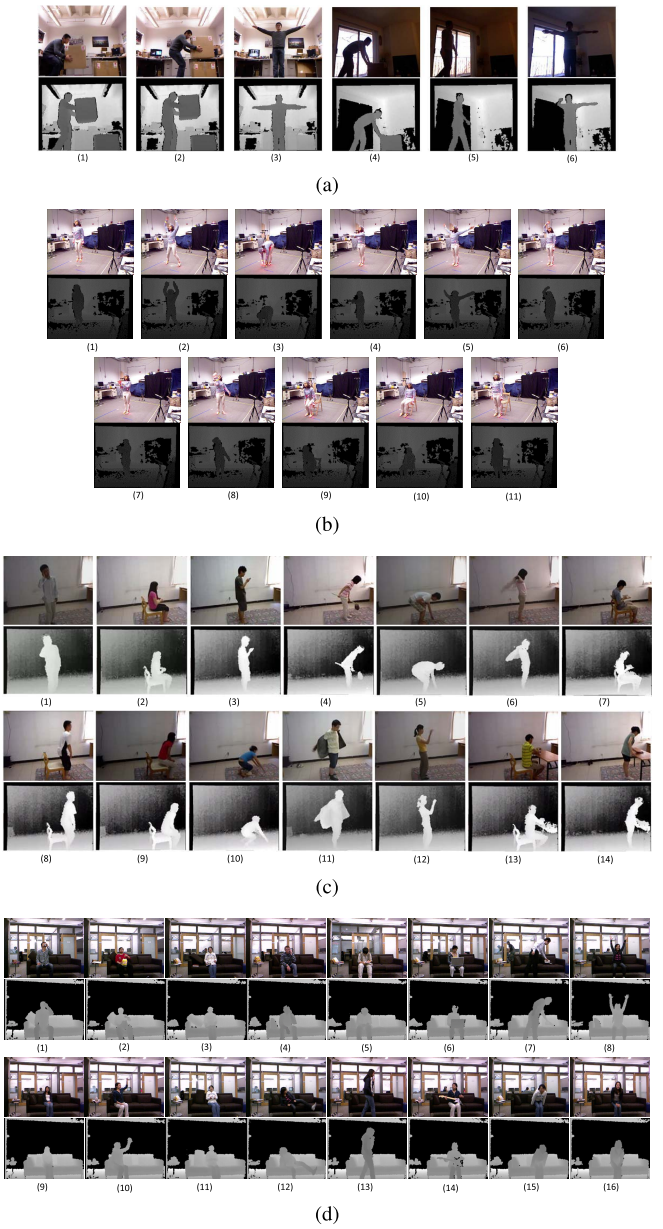


Fig. 9. Exemplary color-depth frames of human activities contained in UTK Action3-D, Berkeley MHAD, ACT4[2], and MSR daily activity 3-D data sets that are collected using color-depth cameras (e.g., Kinect). (a) UTK Action3-D data set contains six human activities. 1: Lifting box. 2: Removing box. 3: Waving. 4: Pushing box. 5: Walking. 6: Signaling. (b) ACT4[2] data set contains 14 human activities. 1: Collapsing. 2: Drinking. 3: Making phone calls. 4: Mopping floor. 5: Picking up. 6: Putting on. 7: Reading book. 8: Sitting down. 9: Sitting up. 10: Stumbling. 11: Taking off. 12: Throwing away. 13: Twisting open. 14: Wiping clean. (c) Berkeley MHAD data set contains eleven activities. 1: Jumping in place. 2: Jumping jacks. 3: Bending. 4: Punching. 5: Two-hand waving. 6: One-hand waving. 7: Clapping hands. 8: Throwing a ball. 9: Sitting down then standing up. 10: Sitting down. 11: Standing up. (d) MSR daily activity 3-D data set contains 16 human activities. 1: Drink. 2: Eat. 3: Read book. 4: Call cell phone. 5: Write on a paper. 6: Use laptop. 7: Use vacuum cleaner. 8: Cheer up. 9: Sit still. 10: Toss paper. 11: Play game. 12: Lie down on sofa. 13: Walk. 14: Play guitar. 15: Stand up. 16: Sit down.

We employ the front-view Kinect data to evaluate our features in this paper, which were captured with a resolution of 640 × 480 at a frame rate of 30 frames/s. Fig. 9(b) shows snapshots of the activities in the Berkeley MHAD data set.

TABLE I

CONFUSION MATRIX OBTAINED BY OUR CoDe4D FEATURES OVER THE UTK ACTION3-D DATA SET. EACH COLUMN CORRESPONDS TO THE PREDICTED CATEGORY AND EACH ROW CORRESPONDS TO THE GROUND TRUTH CATEGORY

	Lifting	Removing	Waving	Pushing	Walking	Signaling
Lifting	88.1	11.9				
Removing	13.4	86.6				
Waving			100			
Pushing	2.1	0.9		97.0		
Walking				5.7	94.3	
Signaling			2.6			97.4

3) *ACT4[2]:* Data set[4] [33] is a large-scale multi-Kinect human activity data set that contains 14 activities performed by 24 subjects in 6844 color-depth sequences. This RGB-D data set was collected in a typical living room environment and has a focus on human daily activities. The color-depth data set obtained from camera 4 is used, which shows side views of the human activities. The data set was captured using a Kinect sensor with a resolution of 640 × 480 and a frame rate of 30 frames/s. Examples of each daily activity from the data set are shown in Fig. 9(c).
4) *MSR Daily Activity 3-D:* Data set[5] [47] contains 16 human activities performed by 10 subjects in 320 color-depth sequences. Each subject performs each activity twice in a standing or sitting position in typical office environments. The color frames have a resolution of 640 × 320, while their respective depth frames have a 320 × 160 resolution. Exemplary color and depth frames of each daily activity from the data set are shown in Fig. 9(d).

C. Activity Recognition Evaluation

Using the above mentioned benchmark color-depth activity data sets, we evaluate the performance of our CoDe4D features, combined with the BoF representation and SVM classifier, on activity recognition. In addition, we compare our system with state-of-the-art activity recognition methods based on the BoF model using color-depth visual data.

1) UTK Action3-D: In this experiment, the data set is divided into training and testing sets: the training data set contains 22 color-depth instances; the remaining 11 instances are used for testing. We adopt accuracy as our measure to evaluate our system's recognition performance.

The confusion matrix obtained by our activity recognition system based on CoDe4D LST features is presented in Table I. It can be observed that our algorithm is able to accurately recognize human activities from color-depth visual data. There are several important phenomena that are worth noting. First, our CoDe4D LST feature is capable of encoding time information, which is indicated by the successful separation

[4]The ACT4[2] data set is available at http://vipl.ict.ac.cn/rgbd-action-dataset/action.

[5]The MSR Daily Activity 3-D data set is available at http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc.

TABLE II
ACCURACY COMPARISON OF OUR METHOD WITH BASELINE AND PREVIOUS APPROACHES OVER THE UTK ACTION3-D DATA SET

Feature detector + descriptor + classifier	Precision
CoDe4D + Color cuboid + LDA [5]	77.7%
CoDe4D + Depth cuboid + LDA [5]	85.5%
CoDe4D + Color-depth cuboid + LDA [5]	91.5%
Color cuboid detector and descriptor [13] + SVM	90.9%
Depth cuboid detector and descriptor [13] + SVM	92.6%
Cuboid + Adaptive MCOH + SVM	93.2%
CoDe4D + HOG/HOF + SVM	92.8%
CoDe4D + Adaptive MCOH + SVM	**93.9%**

TABLE III
COMPARISON OF AVERAGE RECOGNITION ACCURACY OVER C-3 COLOR-DEPTH DATA FROM THE BERKELEY MHAD DATA SET

Feature detector + descriptor + classifier	Precision
Depth Harris3D + HOG/HOF + 1-NN [41]	77.4%
Depth Harris3D + Depth HOG/HOF + 3-NN [41]	76.3%
Depth Harris3D + HOG/HOF + SVM [41]	70.0%
Depth Harris3D + HOG/HOF + MKL-SVM [41]	91.2%
Color cuboid detector and descriptor [13] + SVM	90.5%
Depth cuboid detector and descriptor [13] + SVM	88.7%
Cuboid + Adaptive MCOH + SVM	92.1%
CoDe4D + HOG/HOF + SVM	91.7%
CoDe4D + Adaptive MCOH + SVM	**92.4%**

between lifting and removing activities. Because the BoF encoding and the SVM classifier used in our approach are not capable of modeling time, we can infer that the separation between the sequential activities results from our CoDe4D feature. On the other hand, it can also be observed that there exists a large confusion between lifting and removing. This is because LST features generally cannot capture long-term temporal dependences, due to the fact that LST features only incorporate information contained in the support region, which contains only several frames. Second, human activities, such as pushing and walking, in which the subject crosses the entire horizontal view field of the color-depth camera, are often misclassified by several other activities. For instance, the activity pushing is misclassified as lifting and removing. This phenomenon can be partially explained by the observation that these activities share similar atomic motions with other activities; that is, pushing, lifting, and removing contain similar box-holding and body-moving motions. These similar motions can cause overlaps between the feature sets generated by the activities, which often lead to classification errors.

The recognition system using our CoDe4D features obtains an average accuracy of 93.9% over the UTK Action3-D data set. We compare the performance of our system with the baseline methods using the cuboid detector and descriptor [13], which is applied on a sequence of either color or depth images. The comparison results are presented in Table II. It is observed that our CoDe4D LST features improve recognition accuracy by $\sim$1.3% over depth-cuboid features and $\sim$2% over color-cuboid features. In addition, we compare our recognition system with the approaches reported in our previous work [5], which employed cuboid descriptors (extended from [13]) and the latent Dirichlet allocation [50] model to recognize human activities, as presented in Table II. It is observed that by introducing the adaptive MCOH descriptor as well as applying supervised SVM classifiers, the activity recognition approach significantly outperforms our previous approaches on the UTK Action3-D data set. From Table II, we also observe that features incorporating both color and depth information perform much better than methods based only on color or depth cues, which highlights the importance of encoding color and depth cues in LST feature design.

2) Berkeley MHAD: Following the experimental setup in [41], the first seven subjects are adopted for training and the last five subjects for testing. Experiments and comparisons are conducted based on channel three (C-3) of the depth-layered multichannel data, which generally results in superior performance, as demonstrated in the original work [41]. Accuracy is used as evaluation metric to assess human activity recognition performance in this experiment.

We obtain an average activity recognition accuracy of 93.7% over C-3 color-depth data from the Berkeley MHAD data set. We observe similar phenomena as what we obtained from the experiments using the UTK Action3-D data set, including the ability of our CoDe4D LST feature to capture short-term time dependences. We compare our approach with several baseline approaches using the cuboid detector and descriptor [13]. In addition, we compare with state-of-the-art approaches, such as SVMs with multiple Kernel learning [41], which are evaluated using the same color-depth data from the Berkeley MHAD data set. We present our comparison results in Table III. It can be observed that our system, based on the CoDe4D LST features, obtains state-of-the-art accuracy on the C-3 color-depth data from the Berkeley MHAD data set and outperforms the baseline and previous methods.

3) ACT4^2: Following the experimental setup used in [33], eight human subjects are used for training and the remaining for testing; precision is used as our evaluation metric to assess activity recognition performance. Using this experimental setting, we train our activity recognition system using the training set, and evaluate its performance over the testing set.

An average human activity recognition precision of 81.9% is obtained over the ACT4^2 testing data set, using the proposed CoDe4D LST features. Our activity recognition system is able to distinguish sequential activities, including sitting down and standing up, due to our feature's capability of encoding short-term temporal dependences. Table IV presents comparisons of our complete recognition system with baseline algorithms and methods that obtain the previous state-of-the-art performance on the data set. It is observed that the proposed CoDe4D LST features outperform previous LST features on human activity recognition over the ACT4^2 data set.

4) MSR Daily Activity 3-D: We follow the experiment setup used in [9] in our experiments; accuracy is employed as the performance metric. The experimental results over this data set is reported in Table V. An average activity recognition accuracy of 86.0% is obtained, using our CoDe4D

TABLE IV
COMPARISON OF AVERAGE RECOGNITION PRECISION OVER THE ACT4^2 DATA SET

Feature detector + descriptor	Precision
Harris3D + Color-HOG/HOF [33]	64.2 %
Depth layered multi channel STIPs + HOG/HOF [34]	66.3%
Harris3D + Depth-HOG/HOF [33]	74.5%
Harris3D + Comparative coding descriptor [33]	76.2%
Harris3D + Super feature representation [33]	80.5%
Color cuboid detector and descriptor [13]	70.9%
Depth cuboid detector and descriptor [13]	78.8%
Cuboid + Adaptive MCOH	80.4%
CoDe4D + HOG/HOF	79.2%
Our CoDe4D + Adaptive MCOH	**81.9%**

TABLE V
COMPARISON OF AVERAGE HUMAN ACTIVITY RECOGNITION ACCURACY OVER THE MSR DAILY ACTIVITY 3-D DATA SET

Features	Accuracy
Local occupancy pattern (LoP) features [47]	42.5%
Joint position features [47]	68.0%
Harris3D + Depth-HOG/HOF [38]	79.1%
Depth cuboid detector and descriptor [13]	73.6%
Depth cuboid similarity features [9]	83.6%
Actionlet Ensemble [47]	85.8%
Cuboid + Adaptive MCOH	83.4%
CoDe4D detector + HOG/HOF	85.1%
CoDe4D + Adaptive MCOH	**86.0%**

LST features. Comparisons of our approach with baseline algorithms and previous state-of-the-art methods are also presented in Table V. In addition, to separately evaluate the performance of the Code4D detector and adaptive MCOH descriptor, we conduct the following experiments:

1) comparing CoDe4D detectors with benchmark Cuboid detectors, using the same adaptive MCOH descriptor;
2) comparing MCOH descriptors with benchmark HOG/HOF descriptors, using the same CoDe4D detector;
3) comparing adaptive MCOH descriptors with MCOH descriptors, using the same CoDe4D detector.

These comparisons are reported in Table V, which demonstrate that either CoDe4D detectors or adaptive MCOH descriptors can improve activity recognition accuracy; best performance can be achieved by combining both algorithms.

D. Sensitivity Analysis

In this section, we focus on evaluating the sensitivity of our CoDe4D features to a variety of algorithm parameters that are critical for achieving satisfactory human activity performance. In particular, we investigate our CoDe4D detector's parameters, including color-depth mixture weight, depth scale, and number of neighbors that define local maxima in the 3-D saliency map. In addition, we analyze parameters of our MCOH descriptor, including the number of cells used to divide elevation and azimuth angles. Finally, we investigate how human activity recognition performance is affected by vocabulary size and the number of features per instance, when applying our CoDe4D LST features to form the BoF representation. We analyze our CoDe4D LST feature's sensitivity using threefold cross validation over training sets. This learning-evaluation procedure is performed three times, each applying a different subset for validation. When analyzing sensitivity to a specific parameter, other parameters are set to the values as listed in Section VI-A.

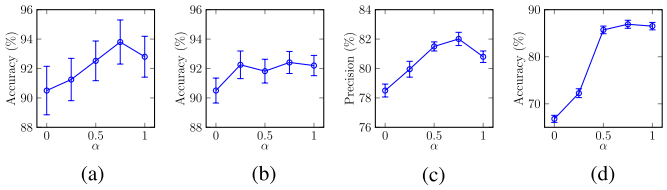


Fig. 10. Sensitivity to the color-depth mixture weight α over different benchmark human activity data sets. Error bars denote deviations resulting from cross variation. (a) UTK. (b) MHAD. (c) ACT4^2. (d) MSR.

1) Color-Depth Mixture Weight: The parameter α is used to balance between color and depth information. Human activity recognition performance over different data sets using different α values is graphically shown in Fig. 10. It is observed that in general, depth cues provide more helpful information than color cues. As shown in Fig. 10, we obtain the best recognition performance when using $\alpha = 0.75$ for the UTK Action3-D and ACT4^2 data sets; when $\alpha \in [0.25, 1]$, we obtain good activity recognition accuracy for the Berkeley MHAD data set; when $\alpha \in [0.75, 1]$, we achieve satisfactory accuracy for MSR Daily Activity 3-D data set.

We observe in our experiments that the choice of the color-depth mixture weight value depends on characteristics of the application. For some data sets (e.g., the UTK Action3-D) that have bad lighting conditions, significant illumination variations, and dynamic background resulting from screens, monitors or TVs, depth information is more important. For other data sets, color information can be weighted more than depth. When the scene is highly cluttered or there exist objects that can absorb IR lights projected by the color-depth camera (as in the Berkeley MHAD data set), depth images generally become very noisy and can contain a large number of black holes with missing depth values and temporally varying shapes. In these cases, using a smaller α to emphasize color information often leads to better recognition performance.

In order to provide an intuitive analysis of how the color-depth mixture weight affects our feature extraction algorithm, we extract CoDe4D LST features from an exemplary instance of the UTK Action3-D, Berkeley MHAD, ACT4^2, and MSR daily activity 3-D data sets. We draw CoDe4D features on an image that fuses two representative intensity frames with their respective depth frames, as shown in Fig. 11. It is observed that using different α values generally results in different sets of STIPs. Moreover, as observed from Fig. 11(b), objects that can absorb IR lights, such as the black cloth in the background, often introduce a large amount of noisy features, which are not relevant to human activities and thus often not helpful to the recognition system. This observation intuitively illustrates why the depth-layered multichannel approach, as used by the original work [41] and this paper, is a necessary component to recognize human activities from the color-depth Berkeley MHAD data set.

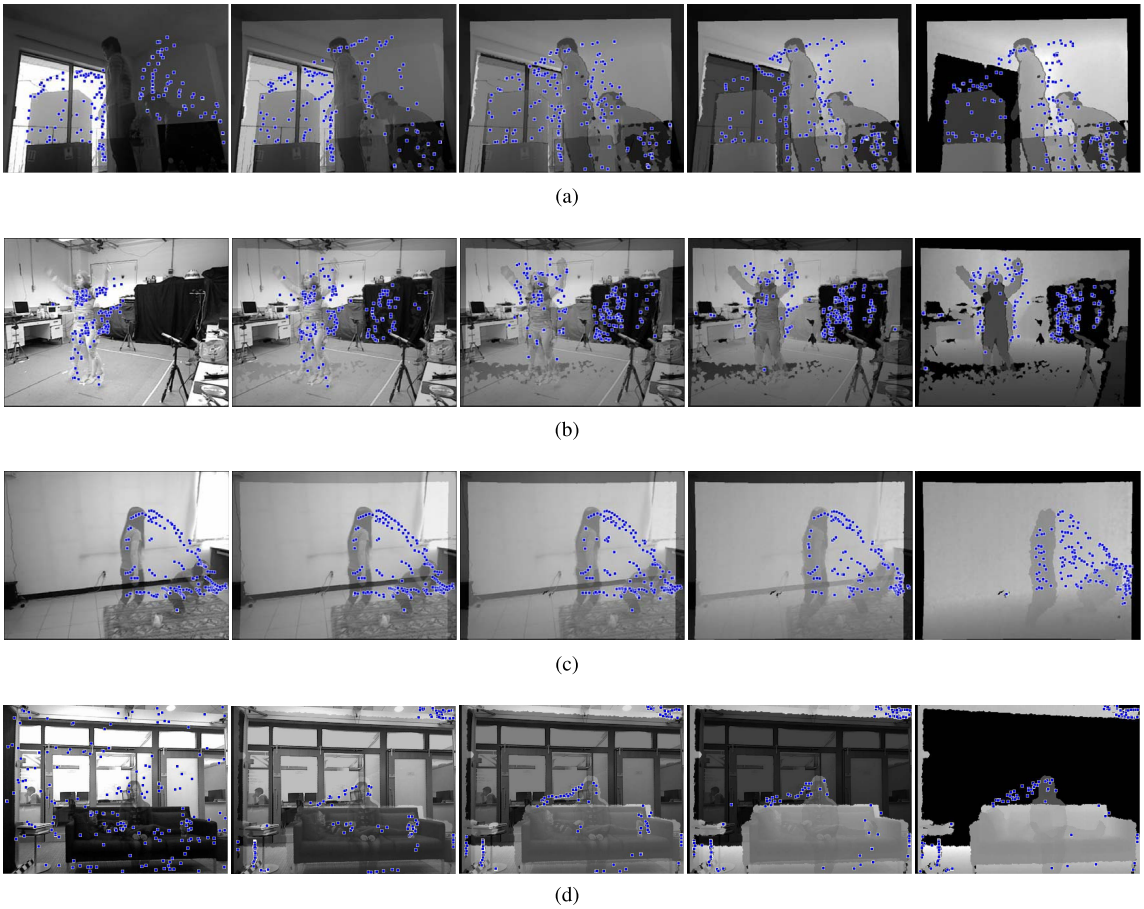


Fig. 11. CoDe4D LST features that are extracted from an instance using different values of color-depth mixture weight, i.e., $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. For a clear display, the detected features are projected onto a 2-D image that is constructed through fusing two representative color and depth frames, using the mixture weight α. A total number of 200 CoDe4D features with largest values in the saliency map are drawn in the figure, using blue boxes with white boundaries. (a) Box-lifting activity from the UTK Action3-D data set. (b) Jumping-jacks activity from the Berkeley MHAD data set (without using depth-layered multichannel data). (c) Stumbling activity from the ACT4^2 data set. (d) Lying on sofa activity from the MSR daily activity 3-D data set.

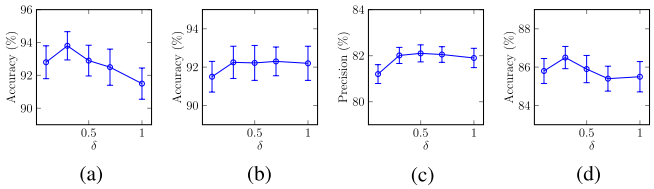


Fig. 12. Sensitivity to the depth scale δ over different benchmark data sets. Error bars denote deviations in cross variation. (a) UTK. (b) MHAD. (c) ACT4^2. (d) MSR.

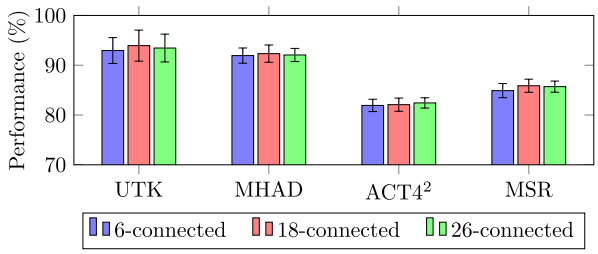


Fig. 13. Sensitivity to the neighborhood connectivity parameter. While accuracy is used as our evaluation measure for UTK Action3-D, Berkeley MHAD, and MSR daily activity 3-D data sets, precision is used for the ACT4^2 data set.

2) Depth Scale: The parameter δ controls the spatial scale along the depth dimension of the cuboid used to detect interest points. Pixels falling outside of the cuboid are not used by our CoDe4D feature detector when building the saliency map. This parameter represents physical distance and is measured using meters. Human activity recognition performance over different data sets using different δ values is graphically shown in Fig. 12. It is observed that when $\delta = 0.3$, our approach generally achieves the best performance over all used data sets. Another interesting observation is that when humans perform activities in open areas as in the Berkeley MHAD and ACT4^2 data sets, our approach is not very sensitive to the depth scale δ. On the other hand, when humans stay close to or interact with other objects as in the UTK Action3-D and MSR daily activity 3-D data sets, a smaller δ is preferred.

3) Neighborhood Connectivity: This parameter defines how to select local maxima from our saliency map in xyt space, which can take values from a finite set $\{6, 18, 26\}$. The activity recognition performance is compared using different numbers of connected neighbors over different data sets. We present our comparison results in Fig. 13. From this figure, we can observe

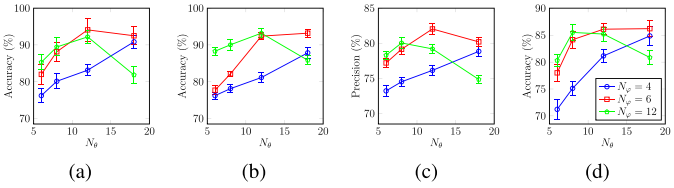


Fig. 14. Sensitivity to the number of bins applied to subdivide the elevation angle φ and the azimuth angle θ. (a) UTK. (b) MHAD. (c) ACT4^2. (d) MSR.

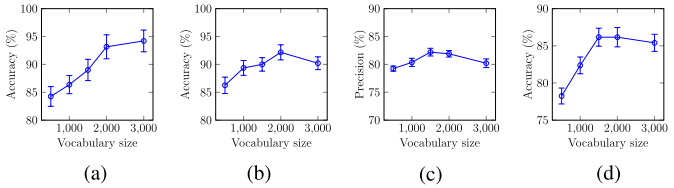


Fig. 15. Variations of human activity recognition performance using different vocabulary sizes. (a) UTK. (b) MHAD. (c) ACT4^2. (d) MSR.

that the performance of our CoDe4D LST features is not very sensitive to the parameter of neighborhood connectivity. This can be partially explained as follows. Although the feature sets obtained using different numbers of connected neighbors can vary, given a fixed number of features to represent an instance, only features with largest values in the saliency map are used. This can lead to similar final feature sets for the data instance and thus result in similar activity recognition performance.

4) Numbers of Angle Bins: These parameters control the granularity of orientation histograms in our MCOH description algorithm. Applying different numbers of bins N_φ and N_θ on the elevation φ and azimuth θ angles, respectively, we assess our system's recognition performance, as shown in Fig. 14. It can be observed that a moderate number of bins often leads to good activity recognition performance. A very coarse-grained subdivision often results in bad performance, because gradient orientations that are significantly different can be assigned to the same cell; consequently, the descriptor is not sufficiently discriminative. On the other hand, a very large number of bins can also reduce performance, since in this situation each cell is generally assigned with less gradient orientations; as a result, the formed feature vector is more sensitive to noise.

5) Vocabulary Size: This parameter controls the number of visual words obtained by clustering to encode the CoDe4D LST features for activity recognition. Variations of recognition performance using different vocabulary sizes are shown in Fig. 15. It can be observed that a vocabulary that has a moderate size often leads to satisfactory recognition performance. This is because, when a small vocabulary size is adopted, features with different patterns can be incorrectly assigned to the same cluster (i.e., visual word); when a very large number of visual words are used, visual features with similar characteristics can be incorrectly assigned to different clusters.

6) Number of Features Per Instance: This parameter defines the total number of CoDe4D LST features to extract from each instance (e.g., 3-D point cloud sequence or color-depth video).

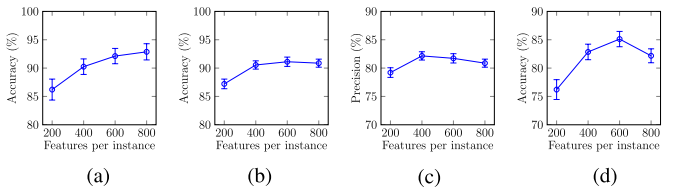


Fig. 16. Variations of human activity recognition performance using different numbers of features per instance. (a) UTK. (b) MHAD. (c) ACT4^2. (d) MSR.

We plot variations of human activity recognition performance using different numbers of features per instance in Fig. 16. It can be observed that extracting 400–600 CoDe4D features to represent each color-depth instance can generally result in satisfactory performance. While using a very small number of features can miss some important local visual cues contained in an instance, extracting a very large number of features can introduce noise, because the low-ranking features can be of poor quality (i.e., weak response in the saliency map).

VII. Conclusion

We introduce a novel LST feature that is able to incorporate both color and depth information contained in a sequence of RGB-D frames. The features are extracted in 4-D space (i.e., $xyzt$), which are able to capture 3-D spatial and 1-D temporal changes of human activities. To detect our CoDe4D features, we apply a 2-D Gaussian filter in the xy dimensions, a pass-though filter in the z dimension, and a Gabor filter in the t dimension. The filtered color-depth information is used to construct a saliency map to encode changes of textures, shapes, and poses. Then, local maxima of the saliency map are selected as our interest points. In order to form a feature vector for each interest point, we propose the MCOH as our feature descriptor to encode spatio-temporal information in the neighborhood of each point. We place a support region, a hyper-cuboid in $xyzt$ space, at each interest point. Then, we compute gradients of the intensity and color patch sequences within the support region, and quantize their orientations using a spherical coordinate-based approach. Our MCOH descriptor incorporates information from both color and depth channels and uses adaptive support region sizes to compensate for linear perspective view changes.

Combining our CoDe4D LST features with BoF representations and SVM classifiers, we construct a complete system to recognize human activities from color-depth visual data. We evaluate the performance of the CoDe4D LST features as well as the complete system using the benchmark UTK Action3-D, Berkeley MHAD, ACT4^2, and MSR daily activity 3-D color-depth activity data sets. Experimental results demonstrate that the proposed CoDe4D LST features present satisfactory representation power and achieve the state-of-the-art activity recognition performance.

References

[1] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[2] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[3] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1993–2008, Nov. 2013.

[4] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in TV shows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2441–2453, Dec. 2012.

[5] H. Zhang and L. E. Parker, "4-dimensional local spatio-temporal features for human activity recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 2044–2049.

[6] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* Jun. 2011, pp. 3273–3280.

[7] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1549–1562, Aug. 2012.

[8] S. Khamis, V. I. Morariu, and L. S. Davis, "A flow model for joint action recognition and identity maintenance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1218–1225.

[9] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2834–2841.

[10] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[12] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2009, pp. 127–137.

[13] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Oct. 2005, pp. 65–72.

[14] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, Sep. 2005.

[15] H. Zhang, W. Zhou, C. Reardon, and L. E. Parker, "Simplex-based 3D spatio-temporal feature description for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2059–2066.

[16] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez, "Selective spatio-temporal interest points," *Comput. Vis. Image Understand.*, vol. 116, no. 3, pp. 396–410, Mar. 2012.

[17] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.

[18] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[19] E. Yu and J. K. Aggarwal, "Human action recognition with extremities as semantic posture representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2009, pp. 1–8.

[20] Z. Zeng and Q. Ji, "Knowledge based activity recognition with dynamic Bayesian network," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 532–546.

[21] A. K. R. Chowdhury and R. Chellappa, "A factorization approach for activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2003, p. 41.

[22] O. Brdiczka, M. Langet, J. Maisonnasse, and J. L. Crowley, "Detecting human behavior models from multimodal observation in a smart home," *IEEE Trans. Autom. Sci. Eng.*, vol. 6, no. 4, pp. 588–597, Oct. 2009.

[23] V. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1896–1909, Dec. 2005.

[24] P. Yan, S. M. Khan, and M. Shah, "Learning 4D action feature models for arbitrary view action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[25] M. Singh, A. Basu, and M. K. Mandal, "Human activity recognition based on silhouette directionality," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 9, pp. 1280–1292, Sep. 2008.

[26] J. Y. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *Proc. AAAI Workshop Pattern, Activity Intent Recognit.*, 2011, pp. 47–55.

[27] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1091–1104, Aug. 2002.

[28] S. Knoop, S. Vacek, and R. Dillmann, "Sensor fusion for 3D human body tracking with an articulated 3D body model," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2006, pp. 1686–1691.

[29] L. Schwarz, D. Mateus, V. Castaneda, and N. Navab, "Manifold learning for ToF-based human body tracking and activity recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 80.1–80.11.

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

[31] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 650–663.

[32] K.-Y. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[33] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2012, pp. 52–61.

[34] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Nov. 2011, pp. 1147–1153.

[35] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Sep. 2007, pp. 357–360.

[36] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 995–1004.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[38] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[39] L. Wang, Y. Qiao, and X. Tang, "Motionlets: Mid-level 3D parts for human motion recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2674–2681.

[40] L. Liu and L. Shao, "Learning discriminative representations from RGB-D video data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1493–1500.

[41] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2013, pp. 53–60.

[42] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 716–723.

[43] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 804–811.

[44] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, Aug. 2007.

[45] V. Bloom, D. Makris, and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2012, pp. 7–12.

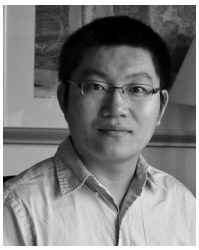
[46] S. R. Fanello, I. Gori, G. Metta, and F. Odone, "Keep it simple and sparse: Real-time action recognition," *J. Mach. Learn. Res.*, vol. 14, pp. 2617–2640, Sep. 2013.

[47] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1290–1297.

[48] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, Sep. 2008.

[49] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 480–492, Mar. 2012.

[50] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.



Hao Zhang (M'15) received the B.S. degree in electrical engineering from University of Science and Technology of China, Hefei, China, in 2006; the M.S. degree in electrical engineering from Chinese Academy of Sciences, Beijing, China, in 2009; and the Ph.D. degree in computer science from University of Tennessee, Knoxville, TN, USA, in 2014.

He is an Assistant Professor with the Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO, USA. His research interests include human-centered robotics, 3-D robotic perception, machine learning, human–robot teaming, and field robotics.

Lynne E. Parker (F'10) received the Ph.D. degree in computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA.

She is currently the Division Director for the Information and Intelligent Systems Division in the Computer and Information Science and Engineering Directorate at the National Science Foundation. While at NSF, she is on leave from the Electrical Engineering and Computer Science Department at the University of Tennessee (UTK), Knoxville, TN, USA, where she is a Professor and previously served as the Associate Department Head. Prior to joining the UTK faculty, she worked for several years as a Distinguished Research and Development Staff Member at the Oak Ridge National Laboratory, Oak Ridge, TN, USA.