Fuzzy Temporal Segmentation and Probabilistic Recognition of Continuous Human Daily Activities

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

***Abstract*—Understanding human activities is an essential capability for intelligent robots to help people in a variety of applications. Humans perform activities in a continuous fashion, and transitions between temporally adjacent activities are gradual. Our Fuzzy Segmentation and Recognition (FuzzySR) algorithm explicitly reasons about gradual transitions between continuous human activities. Our objective is to simultaneously segment a given video into a sequence of events and recognize the activity contained in each event. The algorithm uniformly segments the video into a sequence of non-overlapping blocks, each lasting a short period of time. Then, a multivariable time series is formed by concatenating block-level human activity summaries that are computed using topic models over local spatio-temporal features extracted from each block. Through encoding an event as a fuzzy set with fuzzy boundaries to represent gradual transitions, our approach is capable of segmenting the continuous visual data into a sequence of fuzzy events. By incorporating all block summaries contained in an event, our algorithm determines the activity label for each event. To evaluate performance, we conduct experiments using six datasets. Our algorithm shows promising continuous activity segmentation results on these datasets, and obtains the event-level activity recognition precision of 42.6%, 60.4%, 65.2%, and 78.9% on the Hollywood-2, CAD-60, ACT4^2, and UTK-CAP datasets, respectively.**

***Index Terms*—Human activity recognition, time series segmentation, continuous activities, assistive robotics.**

I. Introduction

At the center of physically assistive robotics applications is how to endow intelligent robots with the capability of interpreting human activities, which is critical for intelligent robots to effectively interact with humans and assist people in human environments. Previous studies [1], [2] in human activity recognition focus on classification of primitive activities contained in short, manually segmented clips, such as walking and hand-waving. However, human activities involve continuous, complicated temporal patterns (for example, grabbing a box then packing and delivering it). Therefore, besides the capability of inferring human activities contained in the segmented events, robots also need the ability to identify the start and end time points of each activity.

Manuscript received March 1, 2014; revised Month Day, Year and Month Day, Year; accepted Month Day, Year. Date of publication Month Day, Year; date of current version Month Day, Month. This paper was recommended by ... of the former IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans (2012 Impact Factor: 2.183).

H. Zhang is with Colorado School of Mines, Golden, CO 80401. (email: hzhang@mines.edu)

W. Zhou and L. E. Parker are with the University of Tennessee, Knoxville, TN 37996. (email: {wzhou4, leparker}@utk.edu)

Color versions of one or more of the figures in this paper are available online at ...

Digital Object Identifier ...

Segmenting and recognizing a sequence of human activities from continuous, unsegmented visual data is more challenging than the task of human activity recognition from a temporally partitioned event that contains a single human activity. Besides difficulties in categorizing human activities in partitioned events, including variations of human appearances and movements, illumination changes, and dynamic backgrounds, recognizing activities in continuous, unsegmented visual data introduces additional challenges. The biggest difficulty of continuous activity segmentation is to deal with the transition effect. Since transitions between temporally adjacent activities occur gradually, their temporal boundaries are vague and even people may not identify when one activity ends and another starts. In addition, generating ground truth to evaluate continuous human activity recognition systems is a challenging task [3]. Errors can arise due to the imprecise activity definition, clock synchronization issues, and limited human reaction time [4]. As a consequence, these challenges result in difficulties in construction of a continuous human activity segmentation and recognition system.

To address this problem, we introduce an algorithm, named *Fuzzy Segmentation and Recognition* (FuzzySR), to temporally partition continuous visual data into a sequence of coherent constituent segments in an unsupervised fashion and to recognize the human activity contained in each individual segment. Our FuzzySR algorithm contains three components (Fig. 1): block-level activity summarization, fuzzy event segmentation, and event-level activity recognition. We employ unsupervised learning since it allows assistive robots to discover new patterns of activities and/or adapt to activity variations of different people. In addition, unsupervised learning takes advantage of the increasing amount of available data perceived by a robot, without the need for human annotation [5], [6].

Our continuous human activity segmentation and recognition algorithm adopts the bag-of-words (BoW) representation based on local spatio-temporal (LST) features [7] that are extracted from visual data. The BoW approach is popular for human activity recognition due to its robustness in real-world environments [8]–[11]. Following the BoW representation, several approaches were proposed to construct a human activity recognition system. Although demonstrated to be effective in recognizing primitive activities in segmented videos [7], [9]–[11], BoW models based on LST features ignore long-term temporal structures of the sequential data, which limits their applications on segmenting continuous visual data that can exhibit temporal patterns. As the BoW model represents videos as a histogram of visual words that are computed from local features, it takes discrete values generally in high

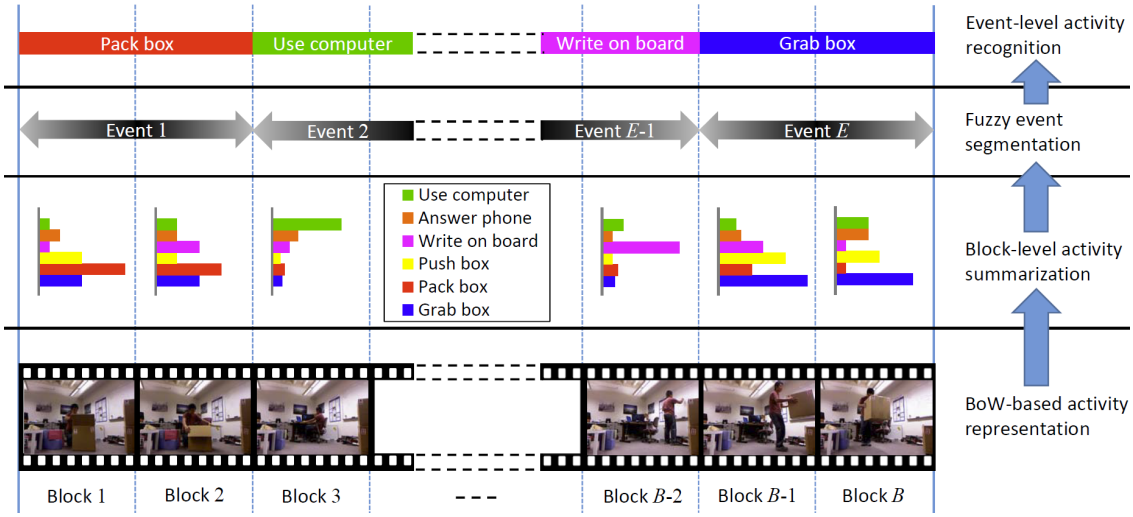


Fig. 1. FuzzySR algorithm for continuous human activity segmentation and recognition. The block-level activity summarization module summarizes the activity distribution of each block by mapping high-dimensional discrete feature space to real-valued activity space. The fuzzy event segmentation module uses the activity summaries to form a multi-variable time series, and uses fuzzy temporal clustering to discover and segment events that are modeled as fuzzy sets. The event-level activity recognition module incorporates summaries of all blocks contained in an event to determine an activity label.

dimensional space, making analysis expensive and generally intractable [12]. Because of this high dimensionality, the BoW model generally cannot be directly used to form a time series to address the problem of temporal pattern analysis.

This paper bridges the divide between temporal activity segmentation and the BoW representation based on LST features [8]. Our approach achieves this objective through applying the *block-level activity summarization*. A *block* is defined as a unit time interval of user-defined duration that contains a short sequence of consecutive video frames, in which the activities performed by a human subject are assumed consistent (i.e., the activities remain the same). As in Fig. 1, our block-level activity summarization partitions a continuous video into a sequence of non-overlapping blocks, and summarizes activity information of each block by mapping the high-dimensional discrete BoW representation in feature space to the real-valued distribution over activities in activity space. Then, the block-level activity distributions are used to form a multi-variable time series. The use of LST features ensures that our algorithm captures the short-term temporal variation within each block.

Another objective is to discover and segment activity events from continuous visual data that can contain a sequence of activities, and to infer an activity label for each individual event. An *event* is defined as a maximum continuous period of time during which the activity label is consistent. Through treating the block-level activity distribution as intermediate information to form a real-valued multi-variable time series, our FuzzySR algorithm follows a fuzzy temporal clustering approach [13] to segment events. We use fuzzy sets to model events and employ fuzzy event boundaries to address gradual transition effects between continuous activities. This procedure is called *fuzzy event segmentation* (Fig. 1). To determine the activity category of a segmented event, we introduce an optimization-based approach that incorporates activity summaries of all blocks contained in the event to make the most appropriate decision. We name this procedure *event-level activity recognition*.

Our work focuses on analyzing temporal characteristics of continuous activities, using temporal fuzzy clustering and unsupervised probabilistic recognition. Our method is a general framework that can work with both color videos and RGB-D visual data. By extending [14], we separate offline training and online testing. To validate our algorithm's effectiveness, we conduct experiments using six benchmark datasets. We use two simple activity datasets to demonstrate how our algorithm segments continuous visual data into events and interprets the activity label in each individual event. Then, we evaluate our algorithm's performance on continuous activity segmentation and recognition. Results indicate promising performance of our method for continuous activity understanding.

The rest of the paper is as follows. After reviewing related work in Section II, we discuss our fuzzy continuous activity segmentation and recognition method in Section III. Results are presented in Section IV. We conclude in Section V.

II. Related Work

A. Human Activity Modeling

Many studies in human activity understanding have focused on recognizing repetitive or punctual activities from short, manually partitioned visual data, which can be acquired from color or color-depth cameras. Instead of discussing supervised learning methods used to classify human activities at the model level, we focus on encoding spatio-temporal information at the feature level to distinguish temporal activity patterns.

A popular space-time representation of human activities is to use centroid trajectories to encode human location variations in visual data. This method, as in [15], encodes a human as a single point, which represents human locations in spatial dimensions. However, this trajectory-based human representation is applicable in the situations when people occupy a small region in an image. Another widely used human activity representation is based on articulated human

body models, such as the skeleton model [16]–[18]. The third category of space-time representations employ a sequence of human shapes [19], including human contours and silhouettes [20], [21], to model temporal activity patterns. Despite the satisfactory recognition performance of the techniques based on body models and human shapes, they depend on human localization and body-part tracking, which involve challenges such as camera motion, occlusion, and dynamic background.

Different from global human representations, local spatio-temporal features have attracted attention, due to the robustness to partial occlusion, slight illumination variation, and image rotation, scaling and translation [8], [22]. Because LST features are directly computed from raw visual data, they can avoid potential failures of preprocessing steps such as human localization and tracking. Dollar et al. [10] detected LST features using separable filters in both spatial and temporal dimensions and described the features using a concatenation-based approach. Laptev et al. [23] detected LST features based on generalized Harris corner detectors, and described these features using a histogram-based method. With the emergence of color-depth cameras, features that are able to incorporate both depth and color information have attracted an increasing attention. [7] introduced the LST feature in 4-dimensional (i.e., $xyzt$) space, which is able to encode both color and depth cues in RGB-D visual data. Xia et al. [24] implemented a feature descriptor based on cuboid similarity to increase the feature's discriminative power.

In this paper, we generally follow this local representation based on LST features to encode human activities. However, we address the task of continuous activity segmentation and recognition in unsegmented sequences. A direct application of LST features to form a time series generally makes the segmentation problem intractable, because the raw LST features can contain a large number of elements in high-dimensional space. We bridge the divide between the continuous activity segmentation problem and the local human representation using LST features by introducing a new layer (i.e., block-level activity summarization) that projects the high-dimensional feature space to the low-dimensional activity space.

B. Temporal Activity Segmentation

Automatic segmentation of complex, continuous activities is important, as intelligent robots deployed in human social environments receive continuous visual data from their onboard perception systems. Without the capability of segmenting the continuous visual data into a temporal sequence of individual activities, it is impossible for robots to understand human behaviors and effectively interact with people.

Previous continuous activity segmentation approaches can be generally grouped into three categories: heuristics, optimization, and change point detection. The first uses simple heuristics to segment human activities from continuous visual data. Fanello et al. [25] calculated a Support Vector Machine (SVM) score from each frame, and then selected the local minima of the score's standard deviation as break points to define the end of a human activity and the start of another. Kozina et al. [26] defined the break points as both local maxima and minima of a given time series. These methods are very sensitive to noise in the time series. When a time series contains multiple variables that usually have a significant amount of noise (Fig. 5(b)), the heuristic methods always over-segment the given continuous visual data, i.e., each activity event is always incorrectly partitioned into a large number of small pieces that may have inconsistent activity labels.

Another framework uses optimization, typically based on discriminative learning, to segment continuous human activities. Shi et al. [27] addressed the human activity segmentation task using a SVM-HMM approach, which is formulated as a regularized optimization problem. A similar approach was introduced by Hoai et al. [28] to jointly segment and classify continuous human activities, which is based on the multi-label SVM-based classification and the discriminative optimization.

The third category is based on change point detection. The earliest and best-known method is the cumulative sum control chart (CUSUM) detector [29], which encodes a time series as piecewise segments of Gaussian means with noise. To process visual data, Zhai et al. [30] applied change point detection to segment video scenes, using heuristic features that are manually defined. Ranganathan [31] performed place classification, using local features such as dense Scale-Invariant Feature Transform (SIFT). Given the satisfactory performance of the methods based on optimization or change point detection, they typically assume fixed boundaries of each activity event, and thus are incapable of modeling gradual transitions between continuous activities in real-world situations.

Different from previous continuous human activity segmentation methods that assume fixed event boundaries [27]–[32], our objective is to explicitly model gradual transitions between temporally adjacent activities. We propose to apply temporal clustering [33] to achieve this objective, which encodes each activity event as a fuzzy set with non-fixed boundaries, instead of segmenting visual data into disjoint events. In addition, the time series used in our algorithm is formulated by concatenating block-level human activity distributions.

There are two research problems different from temporal fuzzy segmentation. The first problem is *fuzzy recognition*, which employs fuzzy methods to recognize activity states, i.e., to assign an activity category to a data instance. For example, Banerjee et al. applied the Gustafson-Kessel [34], [35] or c-means clustering [36] to recognize daily living human activities; Anderson et al. [37], [38] used fuzzy logic based on linguistic antecedent and consequent variables to recognize activity states. The second problem is *background-foreground segmentation*, which aims at localizing humans in the scene and spatially segmenting people from the background. For example, Anderson et al. [39] employed genetic algorithms to segment people and objects out of 3D scenes. Our research partitions continuous data into events along the time dimension using temporal fuzzy clustering. Our probabilistic method also derives fuzzy scores of events in the time dimension; such incrementally changing scores make temporal segmentation and activity recognition results accurate and stable.

III. Fuzzy Segmentation and Recognition

FuzzySR provides a general framework to identify complex, continuous activities from unsegmented visual data with gradual transitions between adjacent activities. The general idea of our algorithm is shown in Fig. 1, and notation appears in Table I. We present how our algorithm is learned during the offline learning phase in an unsupervised fashion (Algorithm 1) and how it is used during the online testing phase (Algorithm 2).

Algorithm 1: Offline unsupervised learning of FuzzySR

Input : K (number of activity clusters), D (dictionary size), $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_B\}$ (a set of blocks)
Output : $\mathcal{M}$ (learned LDA model), $\boldsymbol{D}$ (dictionary)

Extract LST visual features from each block;
Apply k-means method to cluster features into D groups;
Encode each feature using its cluster index (i.e., visual word);
Construct dictionary $\boldsymbol{D}$ that contains all visual words;
Represent each block as a BoW model;
Learn the LDA model $\mathcal{M}$ using the BoW representation (given K) and compute block-level activity distribution;
if *block labels are available* **then**
  Perform semantic mapping using Hungarian method;
end
return $\boldsymbol{D}$ and $\mathcal{M}$

Algorithm 2: FuzzySR for online testing

Input : $\mathcal{W}$ (unsegmented visual data), $\mathcal{M}$ (learned LDA model), $\boldsymbol{D}$ (dictionary)
Output : β (block fuzzy membership), z (event activity category)

Represent $\mathcal{W}$ as a sequence of blocks;
Encode each block as a BoW model, given $\boldsymbol{D}$;
Apply $\mathcal{M}$ on each block to learn activity distribution $\boldsymbol{\theta}$;
Form a multivariate time series using $\boldsymbol{\theta}$ from all blocks;
Compute fuzzy membership β for each block acc. to Eq. (4);
Temporally segment $\mathcal{W}$ into a sequence of events using Eq. (3);
Compute event-level activity assignment z acc. to Eq. (11)
return β and z

A. Block-Level Activity Summarization

The goal of block-level activity summarization is to reduce the input dimensionality in order to form a manageable time series, which is achieved by projecting the high-dimensional feature space to a low-dimensional activity distribution space. Fig. 2 overviews the block-level activity summarization. Our approach is based on LST features (e.g., HOG features for color videos and 4D-LST features for RGB-D data, as specified in Section IV). To construct the dictionary, in the training phase, our approach uses the k-means algorithm to group the LST features (each is a vector containing real values) extracted from training blocks into a given number of clusters. Then each feature vector is encoded by the discrete index of the cluster (referred to as a dictionary word). The dictionary is defined as the collection of all the cluster indices. Given this dictionary, each block can be encoded by a BoW representation (Fig. 2).

Input to our FuzzySR algorithm is an unsegmented video with each frame encoded using the BoW representation based on LST features. This input video $\mathcal{W}$ is temporally partitioned into a sequence of disjoint blocks that have equal length: $\mathcal{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_B\}$, where B is the number of blocks. Each block $\boldsymbol{w}_j$, $j = 1 \ldots B$, is a short sequence of frames. Given a dictionary $\boldsymbol{D}$, which typically has a high dimensionality, each block is represented as a set of discrete visual words that are computed from the LST features using $\boldsymbol{D}$.

TABLE I
Notation for FuzzySR algorithm

Variable	Notation
$\mathcal{W}$	Input unsegmented visual data
$\mathcal{M}$	Learned LDA model
$\boldsymbol{w}$	Block (i.e., a short sequence of frames)
$\boldsymbol{\theta}$	Per-block activity distribution (Eq. (1))
$\boldsymbol{\Theta}$	Time series of $\boldsymbol{\theta}$
$\boldsymbol{l}$	Labels of training blocks
$\boldsymbol{c}$	Activity clusters
$\boldsymbol{D}$	Dictionary
$\boldsymbol{e}(t_s, t_e)$	Event that starts at t_s and ends at t_e
$A_i(t_j)$	Gaussian membership of $\boldsymbol{w}_j$ in $\boldsymbol{e}_i$ (Eq. (5))
$\beta_i(t_j)$	Fuzzy membership of $\boldsymbol{w}_j$ in $\boldsymbol{e}_i$ (Eq. (4))
y_j	Fixed membership of $\boldsymbol{w}_j$ (Eq. (10))
z_i	Activity label of $\boldsymbol{e}_i$ (Eq. (11))
B	Number of blocks in $\mathcal{W}$
E	Number of events in $\mathcal{W}$
K	Number of activity categories
D	Dictionary size
i, j, k	Index of event, block, and activity, respectively

Our algorithm applies a statistical topic model, i.e., Latent Dirichlet Allocation (LDA) [12], to summarize human activity information contained in each block. Given a block $\boldsymbol{w}$, LDA represents each of K activities as the multinomial distribution of all possible visual words in the dictionary $\boldsymbol{D}$. This distribution is parameterized by $\boldsymbol{\varphi} = \{\varphi_{w_1}, \ldots, \varphi_{w_{|\boldsymbol{D}|}}\}$, where φ_w is the probability that the word w is generated by the activity. LDA also models each block $\boldsymbol{w} \subset \mathcal{W}$ as a collection of the visual words, and assumes that each word $w \in \boldsymbol{w}$ is associated with a latent activity assignment z_w. By using the visual words to associate blocks with activities, LDA models a block $\boldsymbol{w}$ as the multinomial distribution over the activities, which is parameterized by $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$, where θ_k is the probability that $\boldsymbol{w}$ is generated by the kth activity. The LDA model is a Bayesian model, which places Dirichlet priors on the multinomial parameters: $\boldsymbol{\varphi} \sim \mathrm{Dir}(\boldsymbol{\beta})$ and $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, where $\boldsymbol{\beta} = \{\beta_{w_1}, \ldots, \beta_{w_{|\boldsymbol{D}|}}\}$ and $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_K\}$ are the concentration hyperparameters.

The objective in block-level activity summarization is to estimate $\boldsymbol{\theta}$, i.e., the per-block activity distribution. However, exact parameter estimation is generally intractable [12]. Gibbs sampling is used to approximately estimate LDA's parameters, which is able to asymptotically approach the correct distribution [40]. When Gibbs sampling converges, the probability of each activity $\theta_k \in \boldsymbol{\theta}$, $k = 1, \ldots, K$, can be estimated by:

$$\theta_k = \frac{n_k + \alpha_k}{\sum_i (n_i + \alpha_i)}, \tag{1}$$

where n_k is the number of times that a word is assigned to the activity $z_w = k$ in the block.

After the per-block activity information is summarized for

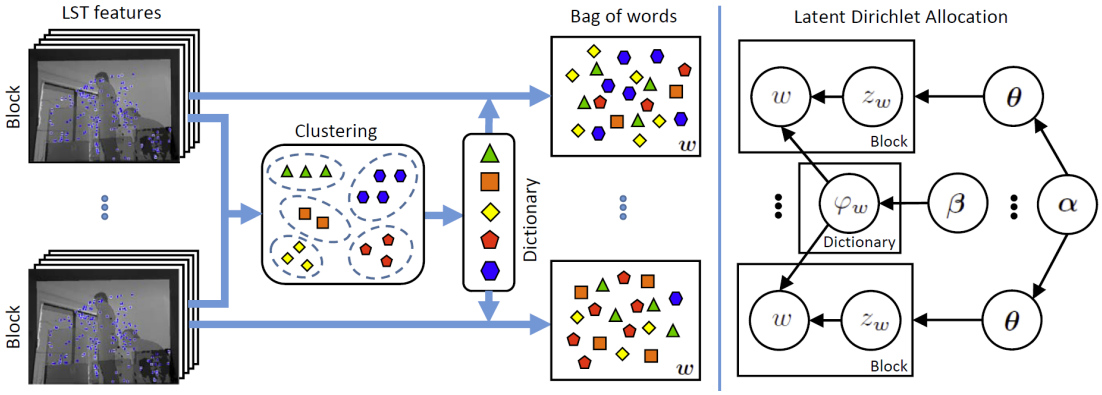


Fig. 2. Block-level activity summarization. After the input continuous visual data are segmented into a sequence of disjoint blocks, the LST features extracted from the frames in each block are converted into discrete visual words using a dictionary. Then, each block is represented by a bag of words $\boldsymbol{w}$, which serves as the input to LDA. The LDA model is used to compute the activity distribution $\boldsymbol{\theta}$ for each $\boldsymbol{w}$. Typically, $\boldsymbol{\theta}$ has a manageable dimensionality that is much lower than the dimensionality of $\boldsymbol{w}$. The human activity summaries from all blocks are applied to form a time series for continuous activity segmentation.

all blocks within the video, a real-valued multi-variable time-series can be formed: $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B\}$, which contains B time-ordered summaries computed at time points $t_1, \ldots, t_B$, where $\boldsymbol{\theta}_j = \{\theta_{j,1}, \ldots, \theta_{j,K}\}^\top$, $j = 1, \ldots, B$, summarizes the activity information contained in the jth block at time t_j. As no ground truth labels are used in the learning process, our per-block activity summarization is performed in an unsupervised fashion.

When semantics (i.e., known activity labels) are available for a subset of blocks (e.g., ground truth of training blocks), the semantics $\boldsymbol{l}$ can be associated with the resulting clusters $\boldsymbol{c}$ obtained by the unsupervised LDA model. For this semantic mapping problem, we use the Hungarian method [41], which finds a bijective (i.e., one-to-one and onto) function $f: \boldsymbol{c} \rightarrow \boldsymbol{l}$ through solving the following:

$$f^\star = \arg\max_{f:\, \boldsymbol{c}\rightarrow\boldsymbol{l}} \sum_{i=1}^{N} \mathbb{1}(\pi_i^l = f(\pi_i^c)). \tag{2}$$

The Hungarian approach [41], [42] formulates this as a bipartite graph matching problem. The graph consists of two sets of nodes (the recognized clusters and the semantic labels) and edge weights are defined as the number of matches.

B. Fuzzy Event Discovery and Segmentation

Given a time series of the block-level activity summaries, the task of continuous human activity segmentation is to seek a sequence of events $\boldsymbol{e}(t_{i-1}, t_i)$, $i = 1, \ldots, E$, where t_i is the temporal boundary of an event that satisfies $t_0 < t_1 < \ldots, < t_E$, and E is the number of events to segment. The segmentation task can be formulated as an optimization problem. Following [13], the optimal event boundaries can be determined through minimizing the sum of the individual event's cost:

$$cost(\boldsymbol{\Theta}) = \sum_{i=1}^{E} \boldsymbol{e}(t_{i-1}, t_i) = \sum_{i=1}^{E} \sum_{j=1}^{B} \beta_i(t_j) \cdot \mathrm{dis}_e(\boldsymbol{\theta}_j, \boldsymbol{v}_i^\theta), \tag{3}$$

where $\mathrm{dis}_e(\boldsymbol{\theta}_j, \boldsymbol{v}_i^\theta)$ denotes the distance between the jth block summary $\boldsymbol{\theta}_j$ and the mean $\boldsymbol{v}_i^\theta$ of $\boldsymbol{\theta}$ in the ith event (i.e., center of the ith cluster), and $\beta_i(t_j)$ denotes the membership of the jth block in the ith event. In [27]–[32], a hard membership is typically used, which satisfies $\beta_i(t_j) = \mathbb{1}(t_i < t_j \leq t_{i+1})$, where $\mathbb{1}(\cdot)$ is the indicator function. However, transitions between temporally consecutive human activities are usually vague. Consequently, changes of the time series formed by the block summaries do not suddenly occur at any particular time point. Thus, it is not practical to define hard event boundaries and not appropriate to model gradual activity transitions using hard memberships.

To address the gradual transition issue, we represent each activity event as a fuzzy set with fuzzy (not fixed) boundaries, and assign the jth block $\boldsymbol{w}_j$ with a fuzzy membership $\beta_i(t_j) \in [0, 1]$ to the ith event $\boldsymbol{e}_i$, as follows:

$$\beta_i(t_j) = \frac{A_i(t_j)}{\sum_{k=1}^{B} A_k(t_j)}, \tag{4}$$

where $A_i(t_j)$ is the Gaussian membership function:

$$A_i(t_j) = \exp\left(-\frac{(t_j - v_i^t)}{2 \cdot (\sigma_i^t)^2}\right), \tag{5}$$

where v_i^t and $(\sigma_i^t)^2$ are the mean and variance of the ith block in the time dimension, respectively. Fig. 3 illustrates modeling events using fuzzy sets with fuzzy boundaries and fuzzy time series segmentation results.

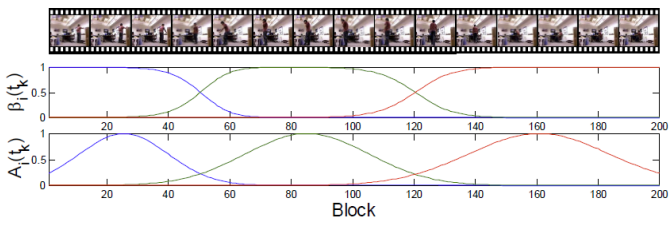


Fig. 3. Modeling events using fuzzy sets that have fuzzy (not fixed) boundaries. A gradual transition always exists between continuous human activities in real-world scenarios. In this example, there exists a transition (block 40–60) between two adjacent activities; another transition (block 105–135) occurs later. By solving Eq. (6), we can obtain the fuzzy segmentation results, which are encoded by the fuzzy membership $\beta(t)$ that is computed using the Gaussian membership function $A(t)$.

To divide a time series of block-level activity summaries into a sequence of events with fuzzy boundaries, we need to estimate the parameters v^t and $(\sigma^t)^2$. A modified Gath-Geva (GG) clustering approach [13], [43] is applied. Through adding time as a variable to each block-level activity summary, i.e., $\boldsymbol{x}=[t, \boldsymbol{\theta}]$, the GG approach favors continuous clusters in time. Assuming that $\boldsymbol{x}$ conforms to the Gaussian distribution, our optimization problem can be defined as follows:

$$\begin{array}{ll} \underset{\boldsymbol{\eta}_i: i=1,\ldots,E}{\text{minimize}} & \sum_{i=1}^{E}\sum_{j=1}^{B}\mu_{i,j}^{m}\,\mathrm{dis}(\boldsymbol{x}_j,\boldsymbol{\eta}_i) \\ \text{subject to} & \sum_{i=1}^{E}\mu_{i,j}=1 \qquad \forall j \\ & 0\le\mu_{i,j}\le 1 \qquad \forall i,j \end{array} \tag{6}$$

where $\mu_{i,j} \in [0,1]$ denotes the membership degree of $\boldsymbol{x}_j$ to the ith cluster parameterized by $\boldsymbol{\eta}_i$, which is computed by:

$$\mu_{i,j}=\frac{1}{\sum_{k=1}^{E}\left(\mathrm{dis}(\boldsymbol{x}_j,\boldsymbol{\eta}_i)/\,\mathrm{dis}(\boldsymbol{x}_j,\boldsymbol{\eta}_k)\right)^{-(m-1)}}, \tag{7}$$

and $m\in(1,\infty)$ denotes the weighting exponent that encodes the fuzziness of the resulting clusters. The weighting exponent [13], [43] of $m=2$ is used here.

The distance $\mathrm{dis}(\boldsymbol{x}_j,\boldsymbol{\eta}_i)$ in Eq. (6) is defined inversely proportional to the probability that $\boldsymbol{x}_j$ belongs to the ith cluster parameterized by $\boldsymbol{\eta}_i$. Since the time variable t is independent of the block summary $\boldsymbol{\theta}$, $\mathrm{dis}(\boldsymbol{x}_j,\boldsymbol{\eta}_i)$ can be factorized:

$$\mathrm{dis}(\boldsymbol{x}_j,\boldsymbol{\eta}_i)=\frac{1}{p(\boldsymbol{x}_j,\boldsymbol{\eta}_i)}=\frac{1}{\alpha_i p(t_j|v_i^t,(\sigma_i^t)^2)p(\boldsymbol{\theta}_j|\boldsymbol{v}_i^\theta,\boldsymbol{\Sigma}_i^\theta)}, \tag{8}$$

where $\alpha_i=p(\boldsymbol{\eta}_i)$ is the prior probability of the ith cluster, which satisfies $\sum_{i=1}^{E}\alpha_i=1$, and t_j and $\boldsymbol{\theta}_j$ in the jth block conform to the Gaussian distribution:

$$\begin{aligned} p(t_j|v_i^t,(\sigma_i^t)^2)&=\mathcal{N}(t_j|v_i^t,(\sigma_i^t)^2)\\ p(\boldsymbol{\theta}_j|\boldsymbol{v}_i^\theta,\boldsymbol{\Sigma}_i^\theta)&=\mathcal{N}(\boldsymbol{\theta}_j;\boldsymbol{v}_i^\theta,\boldsymbol{\Sigma}_i^\theta). \end{aligned}$$

To estimate the parameter, $\boldsymbol{\eta}_i=\{\alpha_i,v_i^t,(\sigma_i^t)^2,\boldsymbol{v}_i^\theta,\boldsymbol{\Sigma}_i^\theta\}$, $i=1,\ldots,E$, the Expectation-Maximization approach is applied to solve Eq. (6), resulting in the following model parameters along the time dimension:

$$v_i^t=\frac{\sum_{j=1}^{B}\mu_{i,j}^m t_j}{\sum_{j=1}^{B}\mu_{i,j}^m},\quad (\sigma_i^t)^2=\frac{\sum_{j=1}^{B}\mu_{i,j}^m(t_j-v_i^t)^2}{\sum_{j=1}^{B}\mu_{i,j}^m}, \tag{9}$$

which can be used to compute the fuzzy membership $\beta_i(t_j)$ of the jth block $\boldsymbol{w}_j$ in the ith event $\boldsymbol{e}_i$, as in Eq. (4). As in Fig. 3, $\beta_i(t_j)$ provides a fuzzy segmentation of the continuous visual data. $\beta_i(t_j)$ can be viewed as the probability that a block belongs to an event: at the gradual transition, the probability of the old activity event decreases, and the probability of the new one increases.

C. Event-Level Activity Recognition

In this paper, the continuous input visual data are uniformly divided into, as well as represented by, a sequence of disjoint blocks. Accordingly, an event can be defined as a maximum sequence of temporally distinct, contiguous blocks that have specific start time, end time, and a consistent human activity label. The objective of event-level activity recognition in our FuzzySR algorithm is to determine these parameters for each event that contains a consistent activity.

To determine the start and end times of an activity event, the computational principle "winner-take-all" is used to represent segmentation results corresponding to the fuzzy membership. Given the fuzzy membership of the jth block, denoted by $\boldsymbol{\beta}_j=[\beta_i(t_j)]$, $i=1,\ldots,E$, its corresponding hard segmentation result y_j can be computed as follows:

$$y_j=\underset{i=1,\ldots,E}{\arg\max}\,\beta_i(t_j) \tag{10}$$

After the hard segmentation result y_j is obtained for each block $\boldsymbol{w}_j$, the human activity label of an event is determined using summaries of all blocks that are contained in the event. Mathematically, given the sequence of block summaries $\boldsymbol{\Theta}=\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_B\}$ and the segmentation results $\boldsymbol{y}=\{y_1,\ldots,y_B\}$, for each event $\boldsymbol{e}_i$, $i=1,\ldots,E$, the activity category z_i can be determined by solving the following optimization problem:

$$z_i=\underset{k=1,\ldots,K}{\arg\max}\;\frac{1}{B}\cdot\sum_{j=1}^{B}\left(\mathbb{1}(y_j=i)\cdot\log\frac{\theta_{j,k}}{\sum_{s=1}^{K}\theta_{j,s}}\right). \tag{11}$$

By computing the probability that the jth block belongs to the kth activity, i.e., $\theta_{j,k}/\sum_{s=1}^{K}\theta_{j,s}$, our algorithm considers the importance of each block under a probabilistic framework to decide the final human activity label of an event. Since topic modeling is used to summarize each block's activity information, $\sum_{s=1}^{K}\theta_{j,s}=1$, $\forall j$ is satisfied. The proposed probabilistic framework could recognize multiple concurrent human activities: when a activity probability threshold is used, activities whose probability are greater than the threshold can be retained (instead of using a max function to select a single activity, as in Eq. (11)).

IV. Empirical Studies

To evaluate our FuzzySR algorithm's performance on segmenting and recognizing continuous human activities, six real-world activity datasets are used. We investigate the performance sensitivity of our algorithm to its parameters, including block size and dictionary size. We chose the benchmark LST features (i.e., HOG features for 2D color videos and 4D-LST features for RGB-D data) to emphasize the performance gain resulting specifically from our temporal fuzzy segmentation and probabilistic recognition approach.

A. KTH Dataset

The KTH dataset contains 600 video sequences captured at 25 frames per second (FPS) with a resolution of 160×120. All videos are recorded using a static camera in a simple environment with homogeneous backgrounds. This dataset contains six human activities: walking, jogging, running, boxing, hand

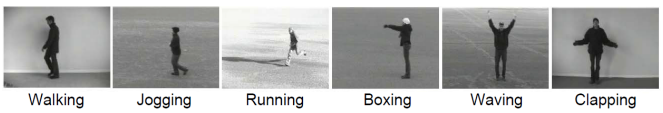


Fig. 4. Representative frames of activities in the KTH dataset.

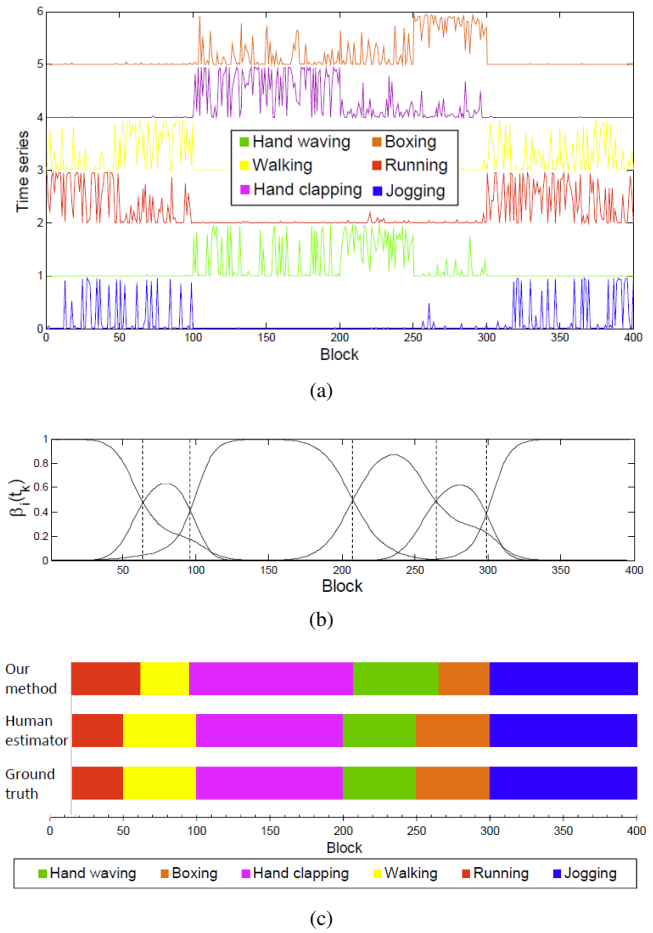


Fig. 5. Results of segmentation and recognition of continuous activities from the KTH dataset. The test video contains six events with instant transitions between human activities. (a) Time series of block-level activity summarizations. (b) Fuzzy segmentation (encoded by the fuzzy membership score $\beta(t)$). (c) Event-level activity recognition results and comparisons with ground truth and results provided by human estimators.

waving, and hand clapping. Each activity is performed by 25 human subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. Representative frames are depicted in Fig. 4.

Since the KTH dataset only contains manually segmented, single-activity videos, to evaluate our FuzzySR's performance on continuous human activity segmentation and recognition, we generate blocks from existing videos in the dataset, and then concatenate these blocks into long videos that contain continuous human activities, as in [28]. We generate 500 blocks, each with a duration of five seconds and 75 frames. We use 100 blocks (12–18 blocks for each activity) to construct an LDA model for block-level summarization, and the remaining 400 blocks for testing using the learned LDA model. As ground truth (i.e., activity label of each data instance) is available, we apply the Hungarian method to associate semantics.

Following [23], we extract low-level LST features through detecting space-time interest points and describing them using histogram of oriented gradients (HOG). Features belonging to the same block are combined. Then, a dictionary of local spatio-temporal words with 400 clusters are constructed using the k-means algorithm. Using this dictionary, features in each block can be converted to visual words. Each block is represented by the BoW model, which serves as the input to our FuzzySR algorithm.

Results with the KTH dataset are in Fig. 5. The time series of the block-level activity summarizations is depicted in Fig. 5(a), which is obtained by applying the learned LDA model on the blocks in the test video. The LDA model is generally capable of summarizing block-level activity information. But activities with upper body movements (e.g., boxing, waving, and hand clapping) are confused with each other. Activities with lower body motions (e.g., walking, jogging, and running) are confused with each other. Jogging and running are not well separated, because these two activities are similar.

Based on the time series of block-level activity summarizations, the fuzzy segmentation result with the KTH dataset appears in Fig. 5(b). Each activity event is encoded by a fuzzy set with fuzzy boundaries. When a current activity is going to transfer to a new activity, the fuzzy membership score $\beta(t)$ of the current activity event decreases and the new event's score increases. Each event obtains its maximum fuzzy membership score at the center of a segment in time dimension, and an activity with a longer event duration generally obtains a more confident segmentation result with a greater fuzzy membership score. These observations indicate our method's effectiveness to model activity transitions and segment continuous activities.

The event-level continuous activity recognition result that is obtained by our algorithm over the KTH dataset is illustrated in Fig. 5(c). Our algorithm's performance is compared with ground truth and results that are manually estimated by human estimators (Fig. 5(c)). Our FuzzySR algorithm well estimates the start and end time points of the events in the test video, and the activity contained in each event is correctly recognized. When the concatenated video is presented to human estimators, they can perfectly identify the events and correctly recognize the activities (Fig. 5(c)).

B. Weizmann Dataset

The Weizmann dataset contains 93 segmented videos with a resolution of 180×144 and is captured at 25 FPS. This dataset is recorded using a static camera in an outdoor environment with a simple background. It contains ten activities performed by nine subjects. The activities include: walking, running, jumping, siding, bending, one-hand waving, two-hands waving, jumping in place, jacking, and skipping. Representative frames are depicted in Fig. 6.

Fig. 6. Exemplary frames of different activities in the Weizmann dataset.

We generate 227 blocks using the existing video clips contained in the Weizmann dataset. Each block has a duration of one second and contains 25 frames. Among the 227 blocks, we generate a test video through concatenating 100 blocks, which contains all ten activities. The test video contains twelve events and each event contains at least five blocks. The remaining blocks are employed to train the LDA model to summarize activity information in each block. We represent each block as a bag of visual words, which are computed by quantizing the HOG features [23] extracted from the block using a dictionary of size 400.

Results with the Weizmann dataset appear in Fig. 7. Our FuzzySR is effective in segmenting a long video that contains continuous activities into fuzzy events; the fuzzy boundaries can well estimate the instant transition between temporally adjacent activities. Fig. 7(b) presents our approach's event-level activity recognition results and comparisons with ground truth and human estimations. Human estimators are able to accurately segment the test video and correctly label the activity contained in each event. Based on the fuzzy event membership score, our FuzzySR achieves comparable segmentation results, and the activity in each event is correctly recognized.

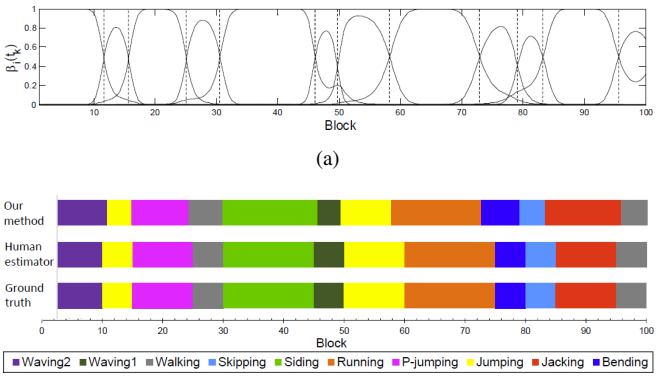


Fig. 7. Results of segmentation and recognition of continuous human activities from the Weizmann dataset. The test video contains twelve events with instant transitions between temporally adjacent activities. (a) Fuzzy segmentation. (b) Event-level activity recognition results and comparisons with ground truth and results provided by human estimators.

C. Hollywood-2 Dataset

The Hollywood-2 dataset [44] is collected from 69 different Hollywood movies with twelve daily activities: answering phone, driving car, eating, fighting person, getting out of car, hand shaking, hugging person, kissing, running, sitting down, sitting up, and standing up. It contains unconstrained activities with challenges including occlusion, camera movement, and lighting changes; different instances of each activity are viewed from different camera angles. Exemplary frames are in Fig. 9.

Following the experimental setup in [44], performance is evaluated using precision; 823 instances are used for training and 884 instances in testing. Following [23], HOG features are used. We randomly generate 500 blocks from instances in the training set, each training block containing 75 frames, which

Fig. 9. Examples of daily activities in the Hollywood-2 dataset, which contain challenges including severe partial occlusions and view point changes.

are used to construct the LDA model and the dictionary that contains 600 visual codewords. We generate 120 testing blocks with the same duration from testing instance. These blocks are used to form a long video that is employed to evaluate our FuzzySR approach's temporal segmentation performance.

Results with the Hollywood-2 dataset are in Fig. 11. Our FuzzySR algorithm can well segment events out of continuous visual data. Recognition errors occur due the significant similarity between the sitting up and standing up activities. We compare our algorithm with unsupervised learning baselines [45], using the same LST features and experimental setups. The baselines unsupervised learning algorithms include partitioning unsupervised learning (e.g., k-means), hierarchical unsupervised learning (e.g., divisive analysis), artificial neural networks (e.g., self-organizing map), and model-based probabilistic unsupervised learning (e.g., mixture of Gaussian and probabilistic latent semantic analysis (PLSA) [46]). The results are presented in Table II. Our algorithm obtains a precision of 42.6%, and outperforms the unsupervised learning baselines, which shows that our approach can well recognize event-level activities, even with the presence of occlusions in the dataset. We also compare our methods with others, which are based on supervised learning (Table II). Supervised learning generally performs better than unsupervised learning since ground truth labels are used in learning to better estimate model parameters.

TABLE II
EVENT-LEVEL AVERAGE RECOGNITION PRECISION WITH THE HOLLYWOOD-2 DATASET.

Approach	Learning	Precision (%)
Marszalek et al. [44]	Supervised	35.5
Derpanis et al. [47]	Supervised	48.0
Gilbert et al. [48]	Supervised	50.9
Wang et al. [49]	Supervised	58.3
Chakraborty et al. [50]	Supervised	58.5
K-means [45]	Unsupervised	29.9
Divisive analysis [45]	Unsupervised	34.8
Self-organizing map [45]	Unsupervised	31.6
Mixture of Gaussian [45]	Unsupervised	30.2
PLSA [46]	Unsupervised	36.7
Our FuzzySR	Unsupervised	42.6

D. CAD-60 Dataset

The CAD-60 dataset [51] contains twelve daily activities: working on computer, brushing teeth, cooking (stirring), cooking (chopping), writing on white board, talking on phone, talking on couch, wearing contact lenses, opening pill container, drinking water, relaxing on couch, and rinsing mouth. These activities are performed by four human subjects in five

Fig. 8. Representative color-depth frames of human activities contained in the CAD-60 dataset that are collected using a Kinect camera in home environments.

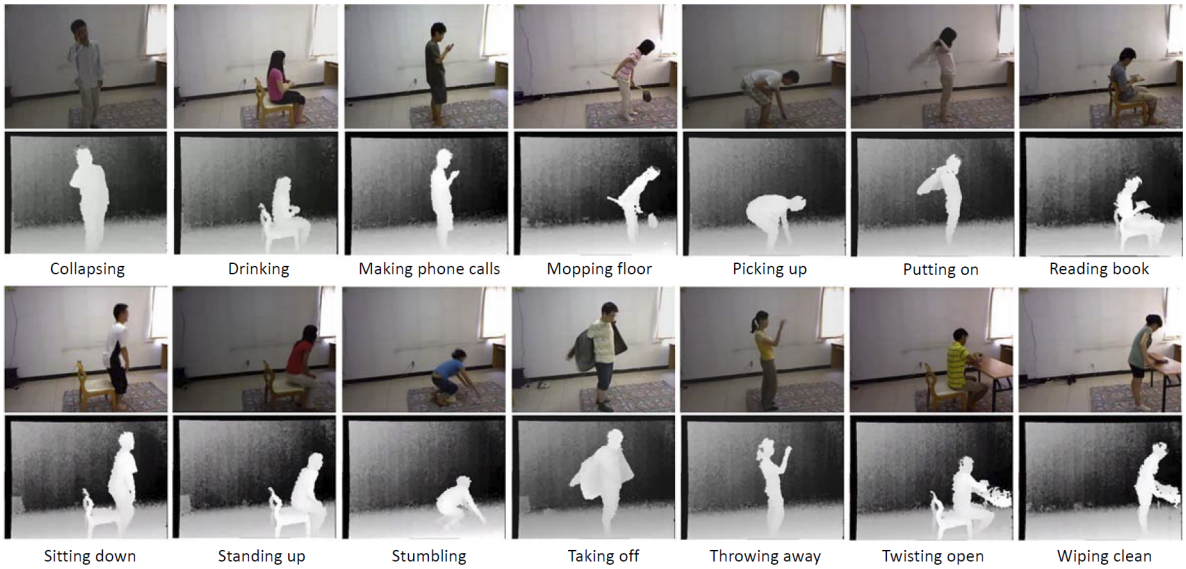


Fig. 10. Exemplary color-depth frames of human activities from the ACT4^2 dataset that are collected using RGB-D cameras in a social environment.

typical human social environments, including office, kitchen, bedroom, bathroom, and living room. The dataset is collected using a Kinect camera; both color and depth frames are provided. Exemplary frames are illustrated in Fig. 8.

Since the CAD-60 dataset only contains manually segmented color-depth videos, each with a single activity, we generate blocks from the dataset frames and then concatenate the blocks to form a long video that contains a sequence of continuous activities, following [28]. As suggested in [51], the system's performance is evaluated according to different locations (e.g., kitchen); in addition, we apply the "new person" experimental setup and use precision and recall as our evaluation metrics. Specifically, we generate a number of 280 blocks, each of which contains 200 color-depth frames, using all instances in the dataset. Then, blocks from one person are used for testing, and blocks from the remaining three persons are used to train a LDA model. Due to their ability to incorporate spatio-temporal color-depth information, we use the 4D-LST features to encode the RGB-D frames, following [7]. A vocabulary that contains 1500 words is constructed and applied to convert a set of visual features from each block to a bag of words.

The continuous human activity segmentation and recognition results in the office scenario are presented in Fig. 12. This scenario includes four activities: working on computer, talking on phone, writing on board, and drinking water. Our FuzzySR obtains satisfactory activity segmentation performance, but recognizing event-level activities from the CAD-60 dataset is challenging, because several activities (e.g., stirring versus chopping, and relaxing on couch versus talking on couch) are similar. We quantitatively evaluate our algorithm's performance on activity recognition at the event level (Table III).

Qualitative evaluation with baseline unsupervised learning

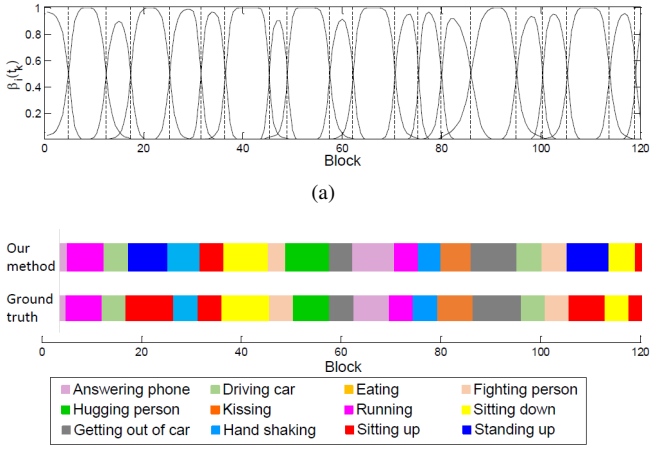


Fig. 11. Results of segmentation and recognition of continuous activities with the Hollywood-2 dataset. The test video contains twenty events with instant transitions between temporally adjacent human activities. (a) Fuzzy segmentation. (b) Event-level activity recognition results and comparisons with ground truth and results provided by human estimators.

algorithms are conducted, using the same features and experimental setups (Table III). Our FuzzySR algorithm obtains superior performance over the baseline unsupervised learning methods. We also compare our unsupervised FuzzySR algorithm with existing supervised methods (Table III). Although supervised learning often outperforms unsupervised learning in the event-level activity recognition task, supervised learning requires ground truth of all instances in the training set to learn model parameters. Since labeling instances is performed manually, it is expensive to obtain ground truth and usually infeasible for a large amount of data in real-world situations.

E. ACT4² Dataset

The ACT4² dataset [55] is a large-scale multi-Kinect dataset that contains 14 activities performed by 24 subjects in 6844

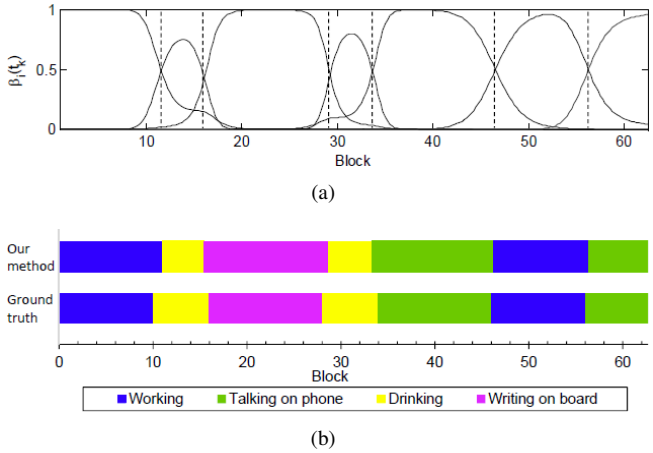


Fig. 12. Results of continuous human activity segmentation and recognition in the office scenario with the CAD-60 dataset. The test sequence contains seven events with instant transitions between temporally adjacent human activities. (a) Fuzzy segmentation. (b) Event-level human activity recognition results and comparison with ground truth.

TABLE III
AVERAGE PRECISION AND RECALL OF EVENT-LEVEL ACTIVITY RECOGNITION OVER THE CAD-60 DATASET.

Approach	Learning	Precision (%)	Recall (%)
Sung et al. [51]	Supervised	67.9	55.5
Koppula et al. [52]	Supervised	80.8	71.4
Ni et al. [53]	Supervised	75.9	69.5
Gupta et al. [54]	Supervised	78.1	75.4
K-means [45]	Unsupervised	48.8	43.1
Divisive analysis [45]	Unsupervised	56.8	51.6
Self-organizing map [45]	Unsupervised	48.9	43.0
Mixture of Gaussian [45]	Unsupervised	51.7	46.2
PLSA [46]	Unsupervised	57.7	55.2
Our FuzzySR	Unsupervised	60.4	55.8

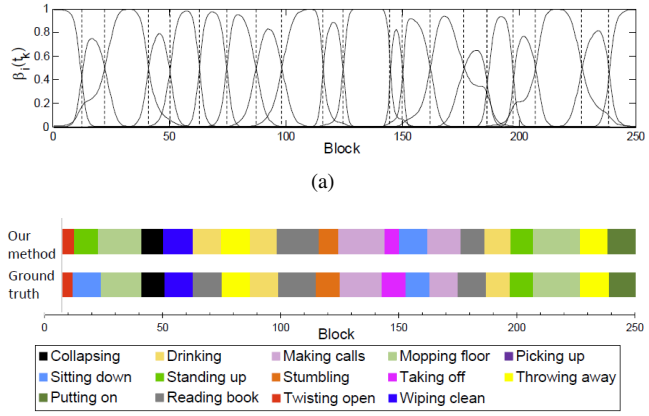


Fig. 13. Results of continuous human activity segmentation and recognition in the office scenario with the ACT4² dataset. The test sequence contains twenty events with instant transitions between temporally adjacent human activities. (a) Fuzzy segmentation. (b) Event-level human activity recognition results and comparison with ground truth.

color-depth sequences. This color-depth dataset is collected in a typical living room scenario and focuses on human activities of daily living. The color-depth data acquired from camera 4 is employed, which shows side views of human activities. The dataset is captured with a resolution of 640×480 and a frame rate of 30 FPS. The color-depth videos are preprocessed by the authors of the dataset, including image smoothing and hole filling. Depth and color frames of each daily activity from the ACT4² dataset are depicted in Fig. 10.

We generate 6000 blocks with each block containing around 100 color-depth frames; the 4D-LST features are employed to encode information from raw color-depth frames, and a vocabulary of size 1500 is used to construct the BoW representation. Following [55], we use blocks from eight humans for training our FuzzySR algorithm and the blocks from the remaining subjects for testing; we apply average precision as the metric to evaluate our FuzzySR's performance on event-level human activity recognition.

Qualitative results with the ACT4² dataset are in Fig. 13. Our FuzzySR algorithm can well segment continuous human activities from the color-depth sequence. However, errors can occur when performing activity recognition from segmented events, due to the strong similarity of several human activities with small motions (e.g., drinking versus reading book). To

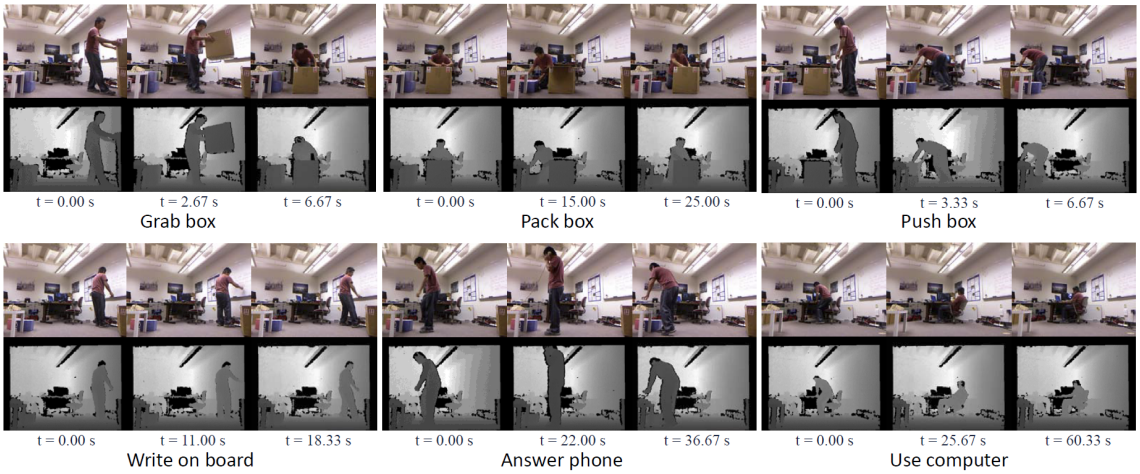
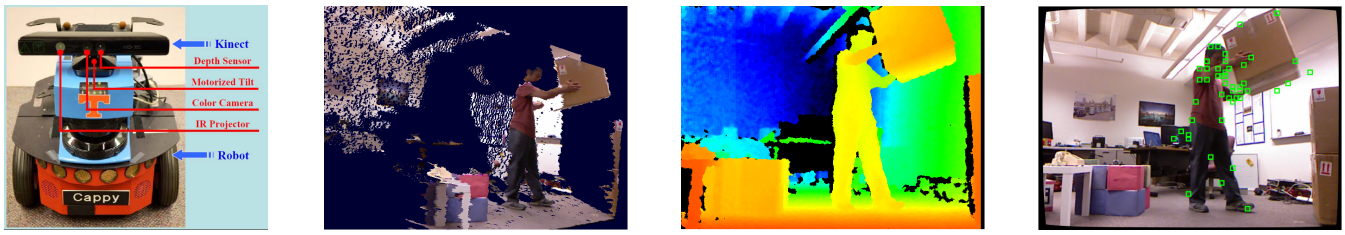


Fig. 14. Typical sequences of the continuous activities in our UTK-CAP dataset. Execution time is labeled under each frame to emphasize the difference in activity durations. In contrast to previous datasets, gradual transitions exist between temporally adjacent activities in our dataset.



(a) (b) (c) (d)

Fig. 15. Setup of our experiments using the newly collected continuous human activity dataset. The Microsoft Kinect color-depth camera is installed on a Pioneer 3DX mobile robot (Fig. 15(a)). Our dataset represents continuous human activities in 3D space (Fig. 15(b)), which contains both depth (Fig. 15(c)) and color (Fig. 15(d)) information. The extracted 4D local spatio-temporal features [7] are also illustrated on the color image (Fig. 15(d)).

TABLE IV
EVENT-LEVEL AVERAGE RECOGNITION PRECISION WITH ACT4^2.

Approach	Learning	Precision (%)
Color-HOGHOF [55]	Supervised	64.2
Depth-HOGHOF [55]	Supervised	74.5
Depth-CCD [55]	Supervised	76.2
DLMC-STIPs [56]	Supervised	66.3
SFR [55]	Supervised	80.5
K-means [45]	Unsupervised	51.5
Divisive analysis [45]	Unsupervised	59.4
Self-organizing map [45]	Unsupervised	53.8
Mixture of Gaussian [45]	Unsupervised	50.9
PLSA [46]	Unsupervised	62.7
Our FuzzySR	Unsupervised	65.2

better understand this error, we perform quantitative evaluation of our FuzzySR algorithm on event-level activity recognition (Table IV). We also compare our algorithm against unsupervised learning baselines and existing supervised approaches. Our algorithm outperforms the unsupervised baselines, and can obtain comparable average event-level recognition precision to several supervised learning approaches (e.g., Color-HOGHOF [55], see Table IV).

F. UTK-CAP Dataset

In real-world scenarios, gradual transitions always exist between temporally adjacent activities. Although the benchmark KTH, Weizmann, CAD-60 and ACT4^2 datasets can be used to generate long sequences, transitions between activities in the concatenated videos occur instantly, which is contradictory to the real-world situation. Accordingly, we employ a continuous activity dataset, i.e., UTK-CAP, to evaluate the effectiveness of our FuzzySR algorithm that explicitly models the gradual transition between adjacent activities in real-life situations.

The UTK-CAP dataset [14] is collected by a Kinect color-depth camera that is installed on a Pioneer 3DX mobile robot (Fig. 15(a)). The dataset contains five color-depth videos. Each video has a duration of around 15 minutes and is recorded at a frame rate of 15 Hz with a resolution of 640×480. Each video contains a sequence of continuous human activities that are performed in a natural way in 3D space. For example, Fig. 15 illustrates the 3D view along with its color and depth images of an activity in the UTK-CAP dataset. The dataset is collected in a small gift store scenario, with a human actor as the store owner performing a sequence of activities related to customer service. An autonomous robot operates to help the human improve productivity. The tasks that the store owner needs to accomplish include posting information and receiving messages on the internet, answering phone calls from customers and suppliers, writing inventory information on a white board, and preparing packages for customers. Six activity categories are designed (Fig. 14):

- Grab box: grab an empty box from the storage area on the right side and bring it to the packing area;

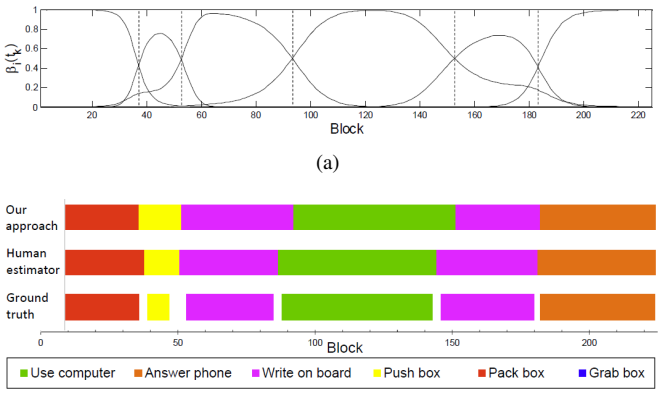


Fig. 16. Results of segmentation and recognition of continuous activities using our continuous activity dataset. The test color-depth sequence contains six events with gradual transitions between temporally adjacent activities. (a) Fuzzy segmentation. (b) Event-level activity recognition results and comparisons with ground truth and results by human estimators. The white spaces in ground truth denote transitions between activities.

- Pack box: put required items into the box in the packing area in the center;
- Push box: push the packed box from the packing area to the delivery area in the far left corner;
- Use computer: operate a computer in the center area;
- Write on board: write notes on a board on the right side;
- Answer phone: answer phone calls on the left side.

We extract 600 blocks, that is 100 blocks for each activity, from the five color-depth videos to learn the LDA model for block-level activity summarization. We represent each block as a bag of visual words, which are computed through quantizing 4D-LST features [7] that are extracted from the blocks using a dictionary of size 400. As an example, the extracted features for the grabbing box activity are shown in Fig. 15(d).

Results over a color-depth video that contains six events are in Fig. 16. The test color-depth video is well segmented by our algorithm, which is able to model gradual transitions between temporally adjacent activities. By representing events as fuzzy sets, our FuzzySR method well estimates the membership of each block. When a block appears in the center of an event, it has a high membership score. If a block approaches the end of the current event, its membership score decreases. Blocks located in gradual transitions have low membership scores for the ongoing event and the new event.

The continuous human activity recognition results over the UTK-CAP dataset are depicted in Fig. 16(b). With the presence of gradual transitions between activities, our FuzzySR approach is still able to correctly recognize continuous activities and well estimate event boundaries. Ground truth is provided by the human actor who performs these activities. Transitions between temporally adjacent activities are explicitly labeled in the ground truth, as denoted by the white spaces in Fig. 16(b). For comparison, five human estimators manually partitioned and recognized the continuous activities contained in the test video. Without knowing the number of activities, human estimators clustered the store owner's activities into 4, 4, 5, 6 and 44 categories, which indicates a strong ambiguity in the definition of the activities in this dataset. Given the number of human activities, human estimators correctly recognized the activities. With the presence of gradual transitions, human evaluators often have difficulty precisely labeling each event's boundaries (Fig. 16(b)). Comparing with human estimations, our FuzzySR algorithm achieves comparable segmentation results over the UTK-CAP dataset (Fig. 16(b)).

TABLE V
EVENT-LEVEL AVERAGE RECOGNITION PRECISION WITH UTK-CAP.

Approach	Learning	Precision (%)
K-means [45]	Unsupervised	67.1
Divisive analysis [45]	Unsupervised	69.2
Self-organizing map [45]	Unsupervised	66.5
Mixture of Gaussian [45]	Unsupervised	68.9
PLSA [46]	Unsupervised	76.2
FuzzySR (based on LDA)	Unsupervised	78.9

In addition, we quantitatively evaluate our FuzzySR algorithm's average recognition precision and compare our result with the unsupervised learning baselines (Table V). Similar to our previous experiments, the FuzzySR algorithm obtains better performance on recognizing event-level human activities.

G. Sensitivity Analysis

We evaluate the sensitivity of our FuzzySR approach to algorithm parameters critical for achieving satisfactory activity segmentation and recognition performance: block size and dictionary size (i.e., number of visual words). In addition, in order to analysis the effect caused by random initialization (as used by the k-means algorithm to construct the dictionary), each set of experiments are performed five times, and an error bar is used to represent the performance variation. Three datasets are employed to perform sensitivity analysis, including Hollywood-2, ACT4^2, and UTK-CAP datasets. When conducting sensitivity analysis to a specific parameter, other parameters are set to the values that are reported in Section IV-C, IV-E and IV-F, for the three used datasets, respectively.

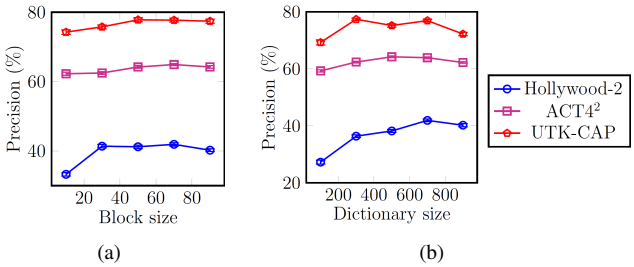


Fig. 17. Our FuzzySR algorithm's sensitivity to the parameters of block size (Fig. 17(a)) and dictionary size (Fig. 17(b)).

1) Block size: This parameter controls the temporal duration of each block (i.e., total number of frames in a block). Performance of event-level activity recognition over different datasets using different block sizes is in Fig. 17(a). A very small block size results in poor event-level activity recognition performance. When fewer frames are contained in the block, the number of extracted visual features is not large enough to represent the activities contained in the block. Block size

cannot be assigned to a very large value, because the block may contain multiple human activities. In real-world applications using cameras with 30 Hz frame rate (e.g., Kinect), using the block size that is in the range between 30 and 60 frames (corresponding to 1–2 seconds) can usually result in satisfactory event-level activity recognition performance.

2) Dictionary size: This parameter controls the number of visual words contained in the dictionary. Since the standard k-means algorithm is employed to construct the dictionary, this parameter also serves as the number of clusters that is provided as a prior to k-means. Event-level activity recognition performance using different dictionary sizes is reported in Fig. 17(b). The dictionary that has a moderate size usually results in satisfactory event-level activity recognition performance. When using a small dictionary size, LST features with different patterns can be incorrectly assigned to the same cluster (i.e., visual word). When a very large dictionary size is employed, visual features with similar characteristics can be incorrectly assigned to different clusters. The dictionary size in the range between 300–800 can achieve good event-level activity recognition results. Our approach is generally not sensitive to different initializations of k-means clustering, demonstrated by the small error bars computed using recognition results in different runs of the experiment.

V. Conclusion

We introduce the FuzzySR algorithm to perform continuous human activity segmentation and recognition. Given a video containing continuous human activities, after uniformly partitioning the video into disjoint blocks, our algorithm computes the human activity distribution of each block through mapping high-dimensional discrete feature space to real-valued activity space. Then, the summaries are used to form a multi-variable time series, and fuzzy temporal clustering is used to segment events. Lastly, our algorithm incorporates all block summaries contained in an event and solves an optimization problem to determine the most appropriate activity label for each event. Our main contributions include explicitly modeling the gradual transition between temporally adjacent human activities, and bridging the divide between the bag-of-word model based on LST features and the continuous human activity segmentation problem. Empirical studies are conducted using six real-world human activity datasets, with a focus on temporally segmenting and probabilistically recognizing continuous human daily activities from both color and RGB-D visual data in human social environments. Results demonstrate our FuzzySR's satisfactory performance, which may allow an autonomous robot to interpret continuous human activities in real-world human social environments.

References

[1] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, pp. 16:1–16:43, Apr. 2011.

[2] P. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, pp. 1993–2008, Nov. 2013.

[3] B. Hard, B. Tversky, and D. Lang, "Making sense of abstract events: Building event schemas," *Memory and Cognition*, vol. 34, pp. 1221–1235, Sept. 2006.

[4] D. Minnen, T. Westeyn, and T. Starner, "Performance metrics and evaluation issues for continuous activity recognition," in *Performance Metrics for Intelligent Systems*, 2006.

[5] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, pp. 299–318, Sept. 2008.

[6] S.-L. Chua, S. Marsland, and H. W. Guesgen, "Unsupervised learning of human behaviours.," in *AAAI Conference on Artificial Intelligence*, 2011.

[7] H. Zhang and L. E. Parker, "4-dimensional local spatio-temporal features for human activity recognition," *IEEE International Conference on Intelligent Robots and Systems*, 2011.

[8] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition.," in *British Machine Vision Association*, 2009.

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

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

[11] Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in *European Conference on Computer Vision*, 2012.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[13] J. Abonyi, B. Feil, S. Nemeth, and P. Arva, "Modified Gath–Geva clustering for fuzzy segmentation of multivariate time-series," *Fuzzy Sets Systems*, vol. 149, pp. 39–56, Jan. 2005.

[14] H. Zhang, W. Zhou, and L. E. Parker, "Fuzzy segmentation and recognition of continuous human activities," in *IEEE International Conference on Robotics and Automation*, In print.

[15] A. K. R. Chowdhury and R. Chellappa, "A factorization approach for activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2003.

[16] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1091–1104, 2002.

[17] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[18] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *IEEE International Conference on Computer Vision*, 2013.

[19] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, pp. 13–24, Jan. 2010.

[20] M. Singh, A. Basu, and M. Mandal, "Human activity recognition based on silhouette directionality," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 9, pp. 1280–1292, 2008.

[21] O. Freifeld, A. Weiss, S. Zuffi, and M. J. Black, "Contour people: A parameterized model of 2D articulated human shape," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, Nov. 2004.

[23] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, pp. 107–123, Sept. 2005.

[24] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

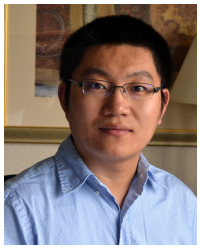
[25] S. R. Fanello, I. Gori, G. Metta, and F. Odone, "Keep it simple and sparse: Real-time action recognition," *Journal of Machine Learning Research*, vol. 14, pp. 2617–2640, Jan. 2013.

[26] M. G. Simon Kozina, Mitja Lustrek, "Dynamic signal segmentation for activity recognition," in *International Joint Conference on Artificial Intelligence*, 2011.

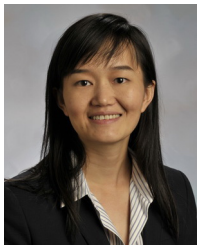
[27] Q. Shi, L. Wang, L. Cheng, and A. Smola, "Discriminative human action segmentation and recognition using semi-Markov model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[28] M. Hoai, Z.-Z. Lan, and F. De la Torre, "Joint segmentation and classification of human actions in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[29] E. S. Page, "Continuous Inspection Schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.
[30] Y. Zhai and M. Shah, "A general framework for temporal video scene segmentation," in *IEEE International Conference on Computer Vision*, 2005.
[31] A. Ranganathan, "PLISS: labeling places using online changepoint detection," *Autonomous Robots*, vol. 32, pp. 351–368, May 2012.
[32] F. Zhou, F. De la Torre, and J. K. Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 582–596, Mar. 2013.
[33] T. Warren Liao, "Clustering of time series data – a survey," *Pattern Recognition*, vol. 38, pp. 1857–1874, Nov. 2005.
[34] T. Banerjee, J. M. Keller, Z. Zhou, M. Skubic, and E. Stone, "Activity segmentation of infrared images using fuzzy clustering techniques," in *World Conference on Soft Computing*, 2010.
[35] T. Banerjee, J. Keller, M. Skubic, and E. Stone, "Day or night activity recognition from video using fuzzy clustering techniques," *IEEE Transactions on Fuzzy Systems*, vol. 22, pp. 483–493, Jun. 2014.
[36] T. Banerjee, J. M. Keller, and M. Skubic, "Resident identification using kinect depth image data and fuzzy clustering techniques," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012.
[37] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M. Aud, "Linguistic summarization of video for fall detection using voxel person and fuzzy logic," *Computer Vision and Image Understanding*, vol. 113, pp. 80–89, Jan. 2009.
[38] D. Anderson, R. Luke, J. Keller, M. Skubic, M. Rantz, and M. Aud, "Modeling human activity from voxel person using fuzzy logic," *IEEE Transactions on Fuzzy Systems*, vol. 17, pp. 39–49, Feb 2009.
[39] D. Anderson, R. Luke, and J. Keller, "Segmentation and linguistic summarization of voxel environments using stereo vision and genetic algorithms," in *IEEE International Conference on Fuzzy Systems*, 2010.
[40] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.
[41] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, vol. 2, pp. 83–97, 1955.
[42] S. M. R.E. Burkard, M. Dell'Amico, *Assignment Problems* (Revised reprint). Society for Industrial and Applied Mathematics, 2012.
[43] I. Gath and A. B. Gev, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 773–780, Jul. 1989.
[44] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[45] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, pp. 264–323, Sept. 1999.
[46] T. Hofmann, "Probabilistic latent semantic indexing," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.
[47] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes, "Action spotting and recognition based on a spatiotemporal orientation analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, p. 527540, 2013.
[48] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[49] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.
[50] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzílez, "S-elective spatio-temporal interest points," *Computer Vision and Image Understanding*, vol. 116, pp. 396–410, Mar. 2012.
[51] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *IEEE International Conference on Intelligent Robots and Systems*, 2012.
[52] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *International Journal of Robotic Research*, vol. 32, no. 8, pp. 951–970, 2013.
[53] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection," *IEEE Transactions on Cybernetics*, vol. 43, pp. 1383–1394, Oct. 2013.
[54] R. Gupta, A. Y.-S. Chia, and D. Rajan, "Human activities recognition using depth images," in *ACM International Conference on Multimedia*, pp. 283–292, 2013.
[55] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *European Conference on Computer Vision Workshops*, 2012.
[56] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *IEEE International Conference on Computer Vision Workshops*, 2011.



Hao Zhang received the Ph.D. in Computer Science from the University of Tennessee, Knoxville in 2014, the M.S. in Electrical Engineering from the Chinese Academy of Sciences in 2009, and the B.S. in Electrical Engineering from the University of Science and Technology of China in 2006. He is an Assistant Professor in the Department of Electrical Engineering and Computer Science at Colorado School of Mines. His research interests include human-robot teaming, robot perception, machine learning, artificial intelligence, and human-centered robotics.



Wenjun Zhou received the Ph.D. from Rutgers, the State University of New Jersey, the M.S. from the University of Michigan-Ann Arbor, and the B.S. from the University of Science and Technology of China. She is an Assistant Professor in the Department of Business Analytics and Statistics, and a faculty affiliate with the Center for Intelligent Systems and Machine Learning at the University of Tennessee, Knoxville. Her research areas are data mining and statistical computing.

Lynne E. Parker received her Ph.D. in computer science from the Massachusetts Institute of Technology. She is 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, Knoxville, where she is Professor and previously served as Associate Department Head. Prior to joining the UTK faculty, she worked for several years as a Distinguished Research and Development Staff Member at Oak Ridge National Laboratory. She is a Fellow of IEEE.