2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Fuzzy Segmentation and Recognition of Continuous Human Activities

Hao Zhang[1], Wenjun Zhou[2] and Lynne E. Parker[1]

Abstract— **Most previous research has focused on classifying single human activities contained in segmented videos. However, in real-world scenarios, human activities are inherently continuous and gradual transitions always exist between temporally adjacent activities. In this paper, we propose a *Fuzzy Segmentation and Recognition* (FuzzySR) algorithm to explicitly model this gradual transition. Our goal is to simultaneously segment a given video into events and recognize the activity contained in each event. Specifically, our algorithm uniformly partitions the video into a sequence of non-overlapping blocks, each of which lasts a short period of time. Then, a multi-variable time series is creatively formed through concatenating the block-level human activity summaries that are computed using topic models over each block's local spatio-temporal features. By representing an event as a fuzzy set that has fuzzy boundaries to model gradual transitions, our algorithm is able to segment the video into a sequence of fuzzy events. By incorporating all block summaries contained in an event, the proposed algorithm determines the most appropriate activity category for each event. We evaluate our algorithm's performance using two real-world benchmark datasets that are widely used in the machine vision community. We also demonstrate our algorithm's effectiveness in important robotics applications, such as intelligent service robotics. For all used datasets, our algorithm achieves promising continuous human activity segmentation and recognition results.**

I. Introduction

In recent decades, human activity recognition has drawn increasing attention from researchers in different fields of study, such as computer vision, ubiquitous computing, machine learning, and robotics [1]–[3]. Most works in human activity recognition focus on simple primitive activities contained in short, manually segmented clips, such as walking and hand-waving, in contrast to the fact that human activities involve continuous, complex temporal patterns, for example, grabbing a box then packing and delivering it.

We are especially interested in peer-to-peer human-robot teaming [3], in which humans and autonomous robots operate collaboratively in the same physical workspace to achieve the same objective. In this application, human activities are always performed in a continuous fashion. Thus, temporal segmentation and recognition of continuous human activities are crucial capabilities for autonomous robots to understand and effectively interact with humans.

Not surprisingly, recognizing a sequence of human activities from a continuous, unsegmented video is considerably more challenging than from a temporally partitioned video that contains a single activity. Besides the well-investigated difficulties to categorize human activities in partitioned clips, such as variations of human appearances and motions, illumination changes and dynamic backgrounds, etc., recognizing human activities in unsegmented visual data poses additional challenges. First, analyzing continuous human activities has to deal with the transition effect, i.e., the transition between temporally adjacent human activities always occurs gradually, and their temporal boundaries are usually vague. Second, humans usually perform multiple activities in parallel. For example, naturally, everyone sits while driving. Third, generating ground truth to evaluate continuous human activity recognition systems is a challenging task. Errors often arise due to clock synchronization issues, limited human reaction time, and imprecise activity definitions [4]. In consequence, these problems result in significant difficulties in construction of continuous activity recognition systems.

To address this important but difficult research problem, we introduce a novel algorithm, named *Fuzzy Segmentation and Recognition* (FuzzySR), to temporally partition a video into coherent constituent segments in an unsupervised fashion and to categorize the activity contained in each individual segment. The main idea of our FuzzySR algorithm is demonstrated in Figure 1, which contains three components: block-level activity summarization, fuzzy event segmentation, and event-level activity recognition.

Our continuous human activity segmentation and recognition algorithm adopts the bag-of-words (BoW) representation based on local spatio-temporal features. The BoW representation is a most popular model for human activity recognition due to its robustness in real-world environments [2], [5]–[7]. Following the BoW representation, several approaches were proposed to construct human activity recognition systems. Although demonstrated to be effective to recognize primitive activities in segmented clips [5]–[8], the BoW model ignores long-term temporal structures of the sequential data, which limits their capability of partitioning continuous videos that exhibit temporal patterns. In addition, since the BoW model encodes videos as a histogram of visual words that are computed from local features, it takes discrete values generally in high dimensional space, which makes analysis directly using the BoW model very expensive and generally intractable [9]. This characteristic limits the BoW model's ability to directly form a time series for temporal pattern analysis.

An important objective of this paper is to bridge the gap between temporal human activity segmentation and the BoW representation, which is not discussed in previous studies to our knowledge. Our approach achieves this objective through applying the *block-level activity summarization*. A *block* is

[1] Hao Zhang and Lynne E. Parker are with the Distributed Intelligence Laboratory, Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Tennessee 37996, USA, {haozhang, leparker}@utk.edu.

[2] Wenjun Zhou is with Department of Statistics, Operations and Management Science, University of Tennessee, Knoxville, Tennessee 37996, USA, wzhou4@utk.edu.

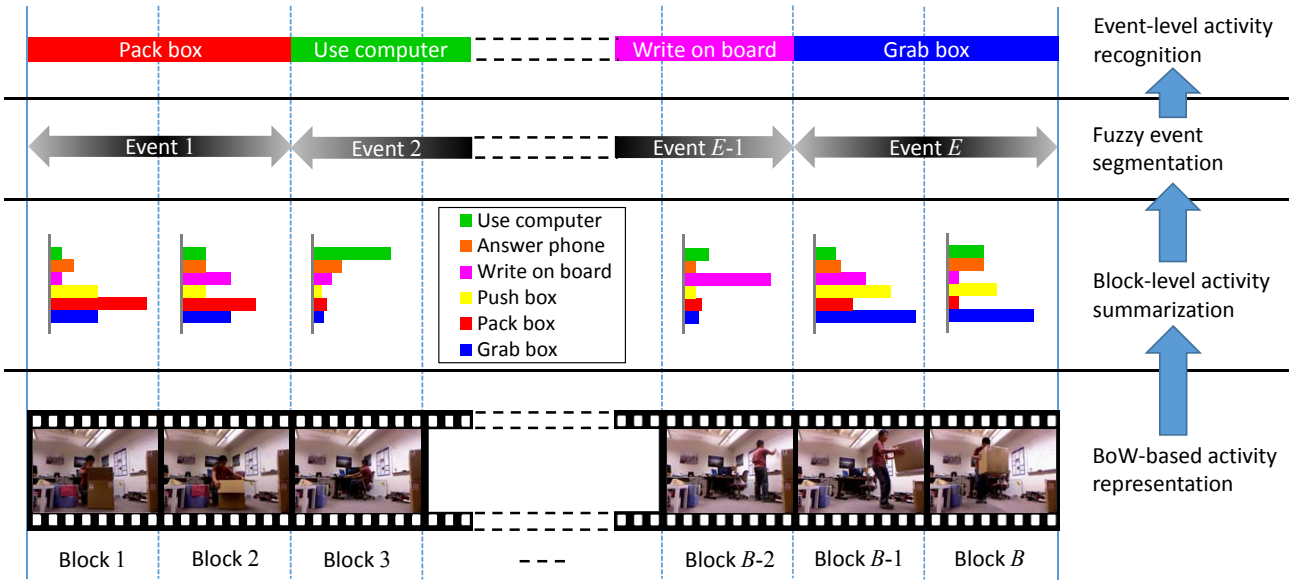


Fig. 1: Illustration of our FuzzySR algorithm for continuous human activity segmentation and recognition. The block-level activity summarization module summarizes activity distribution of each block by mapping high-dimensional discrete feature space to real-valued activity space. The fuzzy event segmentation module uses the summaries to form a multi-variable time series, and applies 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.

defined as a unit time interval of user-defined duration that contains a short sequence of consecutive frames, in which the activity performed by a human subject is assumed consistent. As illustrated in Figure 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. It is noteworthy that the use of local spatio-temporal features also ensures that our algorithm captures temporal variations within each block.

Another objective of this paper is to discover and segment events from a given video that contains a sequence of human 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 constant. Through treating the block-level activity distribution as intermediate information to form a real-valued multi-variable time series, our algorithm follows a fuzzy temporal clustering approach [10] to segment events. We use fuzzy sets to model events and employ fuzzy event boundaries to handle the transition effect. This procedure is called *fuzzy event segmentation*, as depicted in Figure 1. To determine the activity category of a segmented event, we introduce a new, optimization-based method that incorporates summaries of all blocks contained in the event to make the most appropriate decision. We name this method *event-level activity recognition*.

In order to demonstrate the effectiveness of our FuzzySR algorithm, we conduct extensive experiments using benchmark activity datasets that are widely used in the machine vision community. In addition, we introduce a new dataset collected in real-world scenarios that focus on human-robot teaming. It can be difficult to identify the activities contained in the new dataset without discovering temporal patterns and segmenting the events. Our FuzzySR algorithm shows very promising results over the used datasets.

The rest of the paper is structured as follows. After reviewing related work in Section II, we discuss our fuzzy continuous human activity segmentation and recognition approach in Section III. Then, experimental results are presented in Section IV. Finally, we conclude this paper in Section V.

II. Related Work

Segmentation and recognition of continuous human activities is a research topic that involves several key techniques. Previous works relating to this topic are reviewed in detail in this section.

A. Activity Classification

A large number of studies in human activity recognition focused on the problem of recognizing repetitive or punctual activities, such as walking and kicking, from short, manually partitioned visual data such as image sequences and videos, which can be acquired from color or RGB-D cameras.

A most popular methodology to construct a vision-based activity recognition system is based on the BoW model using local spatio-temporal features [2]. Dollar *et al.* [6] extracted such features using separable filters in both spatial and temporal dimensions. Laptev *et al.* [11] detected spatio-temporal features based on generalized Harris corner detectors. Other local spatio-temporal features were also implemented based on the extended Hessian saliency measure [12] and salient

region detectors [13]. Furthermore, with the emergence of color-depth cameras, such as Kinect, several recent works [8] introduced local spatio-temporal features in 4D space, which can incorporate color and depth information in videos.

Temporal structures of single activities contained in mutually segmented videos are also widely investigated in human activity recognition. An early work [14] used Hidden Markov Models (HMMs) to recognize human actions in a sequence of images. Since this work, HMMs have been widely applied to model temporal patterns of human activities [1]. Another popular model to analyze temporal human activity variations is Conditional Random Fields (CRF) [15]. Its extensions, such as hidden CRFs [16], were widely used in recent years. Several space-time activity recognition approaches were also recently introduced based on relationship matching [17], or spatio-temporal phrases [7].

These bag-of-word and temporal models focused on single human activity recognition in partitioned videos. We address a different task: segmentation and recognition of continuous human activities in unsegmented videos.

B. Temporal Activity Segmentation

Automatic segmentation of complex continuous activities is important, since human activities are always continuous in real-world scenarios, which are characterized by a complex temporal composition of single activities. Previous work on temporal segmentation can be generally categorized into two groups: change point detection and temporal clustering.

Change point detection has a long history in statistics and machine learning. The earliest and best-known technique is the CUSUM detector [18], which represents a time series as piecewise segments of Gaussian mean with noise. In recent years, change point detection has drawn increasing attention to process visual data. For example, Zhai et al. [19] proposed to employ change point detection to segment video scenes. Change point detection was also used by Ranganathan [20] to perform place categorization.

Temporal clustering [21] is another popular methodology to segment continuous videos. Especially, several works used temporal clustering to segment visual and motion data into disjoint single-activity events that have hard boundaries [22], [23]. Different from previous methods, we explicitly model gradual transitions between temporally adjacent human activities, following the clustering method proposed by Abonyi et at. [10]. In addition, the time series used in our algorithm is formulated in a new way by computing and concatenating block-level human activity distributions.

III. Our Algorithm

We describe our FuzzySR algorithm for fuzzy continuous human activity segmentation and recognition in this section. FuzzySR provides a general framework to identify complex, continuous human activities from unsegmented videos with gradual activity transitions. In addition, our FuzzySR algorithm bridges the gap between the BoW model and temporal activity segmentation. The idea of our algorithm is presented in Figure 1.

A. Block-Level Activity Summarization

Input to our algorithm is an unsegmented video with each frame encoded using the BoW representation based on local spatio-temporal features. This input video $\mathcal{W}$ is temporally partitioned into a sequence of blocks of 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 set of discrete visual words computed from local spatio-temporal features at time point t_j.

Our algorithm applies a statistical topic model, i.e., Latent Dirichlet Allocation (LDA) [9], to summarize human activity information that is contained in each block. Given a block $\boldsymbol{w}$, LDA represents each of K activities as a 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. By using these visual words to associate blocks with activities, LDA models a block $\boldsymbol{w}$ as a 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 to is to estimate $\boldsymbol{\theta}$, i.e., the per-block activity distribution. However, exact parameter estimation is generally intractable [9]. Gibbs sampling is a widely used technique to approximately estimate LDA's parameters, which is able to asymptotically approach the correct distribution [24]. When Gibbs sampling converges, each activity probability $\theta_k \in \boldsymbol{\theta}$, $k = 1, \ldots, K$, 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 $z = k$ in the block. When trained using labeled data, semantics (i.e., known activity categories) can be associated with the resulting clusters using the Hungarian method [25]. It is noteworthy that, although our discussion is based on the benchmark LDA model, other sophisticated topic models are also directly applicable to our approach.

After the per-block activity information is summarized for 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.

B. Fuzzy Event Discovery and Segmentation

Given a time series of block-level activity summaries, the continuous human activity segmentation task is to seek a sequence of non-overlapping 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.

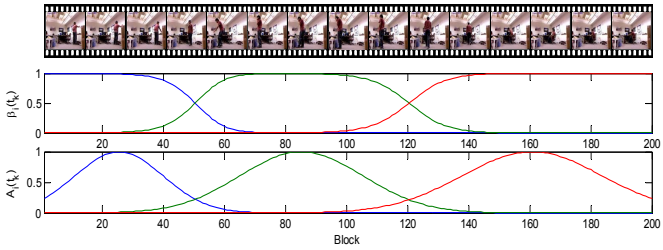


Fig. 2: Illustration of modeling events using fuzzy sets that have fuzzy boundaries. A transition always exists between human activities in real-world scenarios. In the video, there exists an transition (block 40–60) between writing-on-board and answering-phone, and another transition (block 105–135) occurs after it. Through solving the optimization problem in Eq. (5), we can obtain the fuzzy segmentation results, which are encoded by the fuzzy membership $\beta(t)$ that is computed using the Gaussian membership function $A(t)$.

The segmentation task can be formulated as an optimization problem. Following [10], the optimal event boundaries can be determined through minimizing the sum of the individual event's cost:

$$cost(\boldsymbol{\Theta}) = \sum_{i=1}^{E} \boldsymbol{e}(t_{i-1}, t_i) = \sum_{i=1}^{E}\sum_{j=1}^{B} \beta_i(t_j) \cdot \mathrm{dis}(\boldsymbol{\theta}_j, \boldsymbol{v}_i^{\theta}), \quad (2)$$

where $\mathrm{dis}(\boldsymbol{\theta}_j, \boldsymbol{v}_i^{\theta})$ 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. Typically, a hard membership is 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 in the real-world scenario. Consequently, changes of the time series that is formed by block summaries do not suddenly occur at any particular time point. Therefore, it is not practical to define hard boundaries of the events and not appropriate to model gradual activity transition using the hard membership.

To address the gradual transition issue, instead of defining hard event boundaries, we model each event as a fuzzy set with fuzzy boundaries, and assign the jth block with a fuzzy membership $\beta_i(t_j) \in [0, 1]$ to the ith event as follows:

$$\beta_i(t_j) = \frac{A_i(t_j)}{\sum_{k=1}^{B} A_k(t_j)}, \quad (3)$$

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

$$A_i(t_j) = \exp\left(-\frac{(t_j - v_i^t)}{2 \cdot (\sigma_i^t)^2}\right), \quad (4)$$

where v_i^t and $(\sigma_i^t)^2$ are the mean and variance of the ith block in time dimension, respectively. Figure 2 illustrates our idea of modeling events using fuzzy sets with fuzzy boundaries, which also visualizes the fuzzy segmentation results.

To estimate v^t and $(\sigma^t)^2$ in order to divide a time series into a sequence of events with fuzzy boundaries, a modified Gath-Geva (GG) clustering approach [10], [26] is employed. Through adding time as a variable to each block summary, i.e., $\boldsymbol{x} = [t, \boldsymbol{\theta}]$, the GG approach favors continuous clusters in time. Assuming $\boldsymbol{x}$ conforms to the Gaussian distribution, the optimization problem is defined as:

$$\begin{array}{ll} \underset{\boldsymbol{\eta}_i : i=1,\ldots,E}{\text{minimize}} & \sum_{i=1}^{E}\sum_{j=1}^{B} \mu_{i,j}^m \,\mathrm{dis}(\boldsymbol{x}_j, \boldsymbol{\eta}_i) \\ \text{subject to} & \sum_{i=1}^{E} \mu_{i,j} = 1 \qquad \forall j \\ & 0 \leq \mu_{i,j} \leq 1 \qquad \forall i, j \end{array} \quad (5)$$

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:

$$\mu_{i,j} = \frac{1}{\sum_{k=1}^{E} \left(\mathrm{dis}(\boldsymbol{x}_j, \boldsymbol{\eta}_i) / \mathrm{dis}(\boldsymbol{x}_j, \boldsymbol{\eta}_k)\right)^{-(m-1)}}, \quad (6)$$

and $m \in (1, \infty)$ denotes the weighting exponent that encodes the fuzziness of the resulting clusters. A common choice of the weighting exponent [10], [26] is $m = 2$. This value will be used throughout this paper.

The distance function $\mathrm{dis}(\boldsymbol{x}_j, \boldsymbol{\eta}_i)$ in Eq. (5) 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 as:

$$\mathrm{dis}(\boldsymbol{x}_j, \boldsymbol{\eta}_i) = \frac{1}{p(\boldsymbol{x}_j, \boldsymbol{\eta}_i)} = \frac{1}{\alpha_i p(t_j | v_i^t, (\sigma_i^t)^2) p(\boldsymbol{\theta}_j | \boldsymbol{v}_i^{\theta}, \boldsymbol{\Sigma}_i^{\theta})} \quad (7)$$

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:

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

In order to estimate the parameter of each cluster, that is $\boldsymbol{\eta}_i = \{\alpha_i, v_i^t, (\sigma_i^t)^2, \boldsymbol{v}_i^{\theta}, \boldsymbol{\Sigma}_i^{\theta}\}$, $i = 1, \ldots, E$, the Expectation-Maximization approach is applied to solve the optimization problem in Eq. (5), leading to the following model parameter along time dimension:

$$v_i^t = \frac{\sum_{j=1}^{B} \mu_{i,j}^m t_j}{\sum_{j=1}^{B} \mu_{i,j}^m}, \quad (\sigma_i^t)^2 = \frac{\sum_{j=1}^{B} \mu_{i,j}^m (t_j - v_i^t)^2}{\sum_{j=1}^{B} \mu_{i,j}^m}, \quad (8)$$

which can be used to compute the fuzzy membership $\beta_i(t_j)$ of the jth block in the ith event, as defined in Eq. (3).

C. Event-Level Activity Recognition

In this paper, a continuous video is uniformly divided into, as well as represented by, a sequence of blocks. Accordingly, an event can be defined as a maximum sequence of temporally distinct, contiguous blocks having specific start time, end time, and a consistent activity label. The objective of event-level activity recognition is to determine these parameters for each event that contains a consistent activity.

To determine the start time and end time of an event that are also boundaries of an activity, the general computational principle "winner-take-all" is adopted to represent segmentation results corresponding to the fuzzy memberships. Given

the fuzzy membership of the jth block, i.e., $\boldsymbol{\beta}_j = [\beta_i(t_j)]$, $i = 1, \ldots, E$, its segmentation result y_j is computed by:

$$y_j = \arg\max_{i=1,\ldots,E} \beta_i(t_j) \tag{9}$$

After the segmentation result is obtained for each block, the 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 segmentation results $\boldsymbol{y} = \{y_1, \ldots, y_B\}$, for each event $\boldsymbol{e}_i$, $i = 1, \ldots, E$, the activity category z_i is determined by solving the following optimization problem:

$$z_i = \arg\max_{k=1,\ldots,K} \sum_{j=1}^{B} \left(\mathbb{1}(y_j = i) \cdot \log \frac{\theta_{j,k}}{\sum_{s=1}^{K} \theta_{j,s}} \right). \tag{10}$$

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 in a probabilistic fashion to decide the final activity label of an event. In our case, since topic modeling is applied to summarize each block's activity information, $\sum_{s=1}^{K} \theta_{j,s} = 1$, $\forall j$ is satisfied.

IV. Empirical Study

This section describes evaluation results of our FuzzySR algorithm to fuzzily segment and classify continuous human activities over three real-world datasets: KTH and Weizmann datasets, and a newly collected dataset containing continuous human activities, which is recorded using a color-depth camera that is installed on a mobile robot. We run FuzzySR on long video sequences to partition events and label each event with an activity class. Then, we compare our FuzzySR algorithm's segmentation and recognition results with ground truth and results provided by human estimators.

A. KTH Dataset

The KTH dataset contains 2391 video sequences that were captured at 25 frames per second (FPS) with a resolution of 160×120. All video sequences were recorded using a static camera in the environment with homogeneous backgrounds. This dataset contains six human activities: walking, jogging, running, boxing, hand 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 of each activity are depicted in Figure 3.

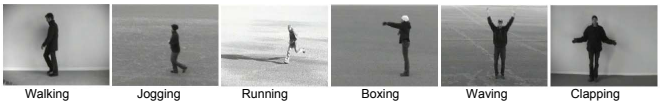
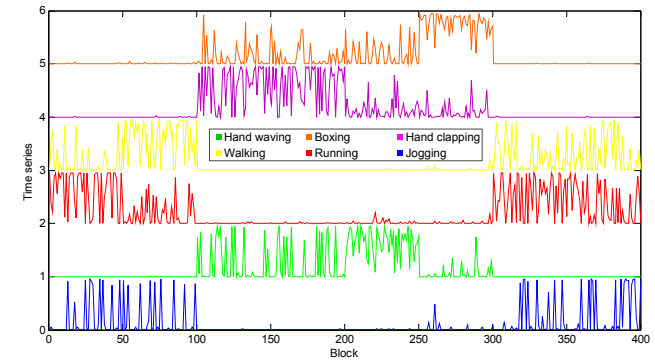
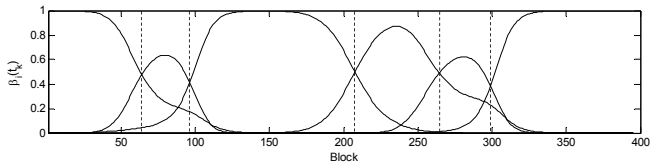


Fig. 3: Representative frames of activities in KTH dataset.

Since the KTH dataset only contains manually segmented, single-activity videos, to evaluate our method's performance of 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



(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.

Fig. 4: Experimental results of segmentation and recognition of continuous activities from the KTH dataset. The test video contains six events with instant transitions between activities.

continuous human activities. Specifically, we generate 500 blocks, each of which has a duration of five seconds and contains 75 frames. We use 100 blocks (around 12–18 blocks for each activity) to construct an LDA model for block-level summarization, and the remaining 400 blocks for testing.

Following [11], we extract local spatio-temporal features through detecting space-time interest points and describing them using histogram of oriented gradients (HOG). Features belonging to the same block are combined together. Then, a dictionary of local spatio-temporal words with 400 clusters are constructed using the k-means quantization. Using this dictionary, features in each block can be converted to visual words. Accordingly, each block is represented by the BoW model, which serves as the input to our algorithm.

Experimental results over the KTH dataset are presented in Figure 4. The time series of block-level human activity summarization is illustrated in Figure 4a, which is obtained by using the learned LDA model on the blocks in the video. This observation demonstrates that the LDA model is capable of summarizing block-level activity information. In addition, it can be observed that activities with upper body motions

(e.g., boxing, waving, and hand clapping) are easily confused with each other. Similarly, activities with lower body motions (e.g., walking, jogging, and running) are confused with each other. Especially, jogging and running are not well separated, because these two activities are extremely similar.

Based on the time series of block-level activity summarizations, the fuzzy segmentation result obtained by our method over the KTH dataset is graphically presented in Figure 4b. It can be observed that each event is encoded by a fuzzy set that has fuzzy boundaries. When a current activity is going to transfer to a new activity, the fuzzy membership score $\beta(t)$ of the current event decreases and simultaneously the new event's score increases. Furthermore, we observe that each event obtains its maximum fuzzy membership score at the center of a segment in the time dimension, and an activity with a longer duration generally obtains a more confident segmentation result with a greater fuzzy membership score. These observations indicate our algorithm'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 shown in Figure 4c. Our algorithm's performance is also compared with ground truth and results that are manually estimated by human estimators, which are depicted in Figure 4c. It can be observed that our 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, due to the clear, instant transition between temporally adjacent activities, human estimators can perfectly recognize the events and correctly classify the activities contained in each event, as presented in Figure 4c.

B. Weizmann Dataset

The Weizmann dataset contains 93 segmented video clips with a resolution of 180×144 and was captured at 25 FPS. This dataset was recorded by a static camera in an outdoor environment with a simple background. The dataset contains ten activities that are performed by nine human subjects. The full activity list is: walking, running, jumping, siding, bending, one-hand waving, two-hands waving, jumping in place, jacking, and skipping. Representative frames showing these activities are depicted in Figure 5.

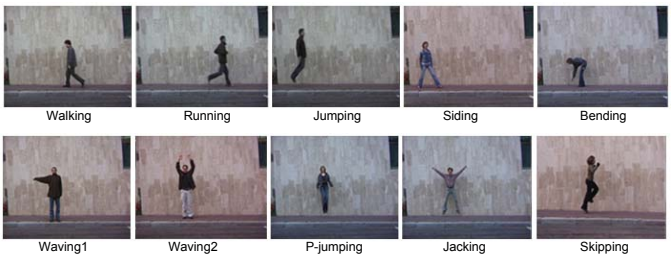
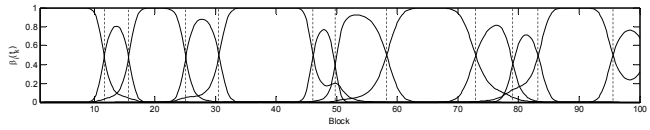
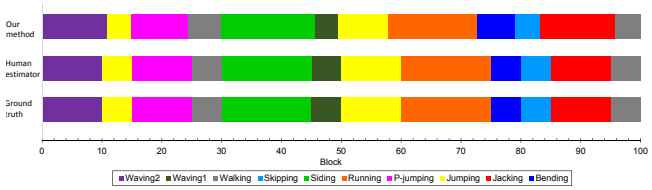


Fig. 5: Exemplary frames of activities in Weizmann dataset.

Similar to our previous experimental settings, 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 by 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 local spatio-temporal features [11] extracted from the block using a dictionary of size 400.



(a) Fuzzy segmentation (encoded by the fuzzy membership score $\beta(t)$).



(b) Event-level activity recognition results and comparisons with ground truth and results provided by human estimators.

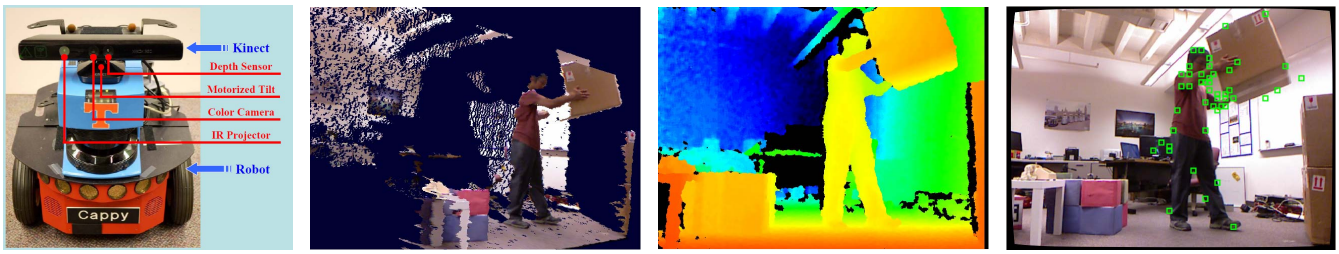
Fig. 6: Experimental results of segmentation and recognition of continuous activities from the Weizmann dataset. The test video contains twelve events with instant transitions between temporally adjacent activities.

Experimental results over the Weizmann dataset are graphically presented in Figure 6. It can be observed from Figure 6a that our method is very effective to segment a long video that contains continuous human activities into fuzzy events; the fuzzy boundaries can well estimate the instant transition between temporally adjacent activities. Figure 6b shows our approach's event-level human activity recognition results and comparisons with ground truth and human estimations. Due to the instant transition between human activities in the test video, human estimators are able to accurately segment the test video and correctly label the activity contained in each event. In addition, it can be observed that, based on the fuzzy event membership score, our FuzzySR achieves comparable segmentation results, and the activity contained in each event is correctly recognized.

C. Continuous Activity Dataset

In real-world scenarios, transitions always exist between temporally adjacent activities. Although the benchmark KTH and Weizmann datasets can be used to generate long videos, activity transitions in the concatenated video are assumed to occur instantly, which is contradictory to the actual situation. Therefore, a continuous activity dataset is needed to demonstrate our algorithm's effectiveness to model gradual activity transitions in real-world scenarios.

To the best of our knowledge, there are no publicly available labeled video datasets adequate to evaluate continuous human activity recognition systems. Due to the lack of such



(a) Camera installation (b) 3D view of an activity (c) Depth Frame (d) 4D-LST features

Fig. 7: 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 (Figure 7a). Our dataset represents continuous human activities in 3D space (Figure 7b), which contains both depth (Figure 7c) and color (Figure 7d) information. The extracted 4D local spatio-temporal features [8] are also illustrated on the color image (Figure 7d).

a dataset, we collect a new one using the Microsoft Kinect color-depth camera that is installed on a Pioneer 3DX mobile robot, as shown in Figure 7a. The dataset contains five color-depth videos. Each video has a duration of around 15 minutes and is recorded at 15 FPS with a resolution of 640×480. Each color-depth video contains a sequence of continuous human activities that are performed in a natural way in 3D space. For example, Figure 7 illustrates the 3D view along with its color and depth images of an activity in our newly collected 3D human activity dataset.

Our dataset is collected in the scenario of a small gift store, in which the human actor plays a role of the store owner and performs a sequence of activities related to customer service. A robot is used to operate in the same environment to help the human improve productivity. During the experiment, the robot is assumed to stay in an observing state without any movements. 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. In this scenario, six activities are designed, as illustrated in Figure 8:

- Grab box: grab an empty box from the storage area on the right side and bring it to the packing area;
- 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, i.e., 100 blocks for each activity, from five long videos to learn the LDA model for block-level activity summarization. We represent each block as a bag of visual words, which are computed by quantizing the 4D local spatio-temporal features [8] extracted from the block using a dictionary of size 400. The 4D features encode variations in spatial and temporal dimensions, and incorporate both color and depth information. For example, the extracted features for the grabbing box activity are shown in Figure 7d.

Experimental results over a color-depth video that contains

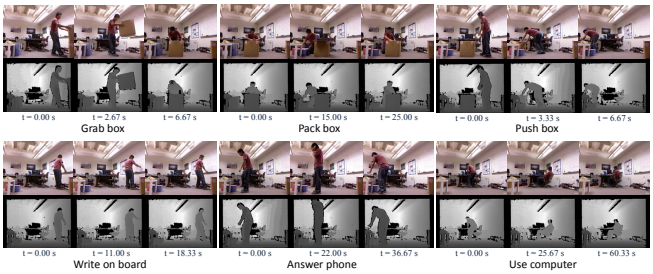
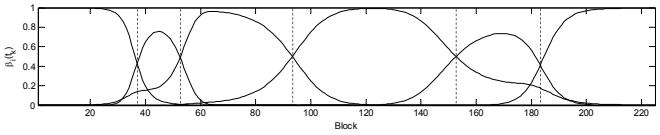


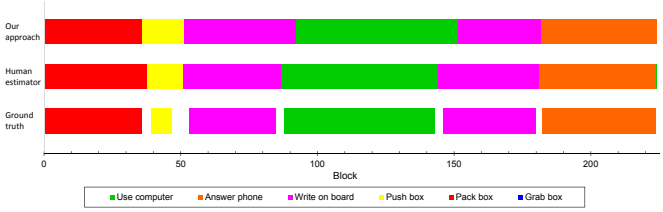
Fig. 8: Typical sequences of the continuous human activities in our dataset. Execution time is labeled under each frame to emphasize the difference in activity durations. In contrast to previous datasets, gradual transitions exist between temporally adjacent activities in our dataset.

six events are depicted in Figure 9. It can be observed from Figure 9a that the test color-depth video is well segmented by our algorithm, which is able to model gradual transitions between temporally adjacent activities. By encoding events as fuzzy sets, our algorithm 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 to the end of the current event, its membership score decreases. Blocks located in transitions generally have low membership scores for the ongoing event and the new event.

The continuous human activity recognition result over our dataset is presented in Figure 9b. It can be observed that, with the presence of gradual transitions between activities, our FuzzySR algorithm is still able to correctly recognize continuous activities and well estimate event boundaries. In this experiment, 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 shown in Figure 9b. For comparison, we invited five human estimators to segment and recognize the continuous activities contained in the test video. Without knowing the number of activities, human estimators clustered the activities into 4, 4, 5, 6 and 44 categories, which indicates a strong ambiguity on the definitions of the activities in our dataset. Given the number of activities, human estimators correctly recognized the activities. On the other hand, with



(a) Fuzzy segmentation (encoded by the fuzzy membership score $\beta(t)$).



(b) Event-level activity recognition results and comparisons with ground truth and results provided by human estimators. The white spaces in the ground truth denote transitions between activities.

Fig. 9: Experimental results of segmentation and recognition of continuous activities using our continuous activity dataset. The test color-depth video contains six events with gradual transitions between temporally adjacent activities.

the presence of gradual transitions, human evaluators may have difficulty precisely labeling each event's boundaries. These phenomena can be seen in Figure 9b. Comparing with human estimations, our FuzzySR algorithm achieves comparable segmentation and recognition results over this activity dataset, as demonstrated in Figure 9b.

V. Summary

We propose the FuzzySR algorithm to perform continuous human activity segmentation and recognition. Given a video containing continuous activities, after uniformly partitioning the video into blocks, our algorithm computes the 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 are twofold:

- We bridge the gap between bag-of-word models based on local spatio-temporal features and continuous human activity segmentation problems;
- We explicitly model gradual transitions between temporally adjacent human activities.

Extensive experiments using real-world datasets demonstrate our FuzzySR algorithm's satisfactory performance on continuous activity segmentation and recognition, which can allow an autonomous robot to interpret human activities in real-world scenarios.

References

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

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

[3] J. R. Hoare and L. E. Parker, "Using on-line conditional random fields to determine human intent for peer-to-peer human robot teaming," in *IEEE International Conference on Intelligent Robots and Systems*, 2010.

[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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European Conference on Computer Vision*, 2008.

[13] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, pp. 710–719, Jun. 2005.

[14] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1992.

[15] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning*, 2001.

[16] S. B. Wang, A. Quattoni, L.-P. Morency, and D. Demirdjian, "Hidden conditional random fields for gesture recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[17] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *IEEE International Conference on Computer Vision*, 2009.

[18] E. S. Page, "Continuous Inspection Schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[19] Y. Zhai and M. Shah, "A general framework for temporal video scene segmentation," in *IEEE International Conference on Computer Vision*, 2005.

[20] A. Ranganathan, "PLISS: labeling places using online changepoint detection," *Autonomous Robots*, vol. 32, pp. 351–368, May 2012.

[21] T. Warren Liao, "Clustering of time series data – a survey," *Pattern Recognition*, vol. 38, pp. 1857–1874, Nov. 2005.

[22] 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.

[23] 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.

[24] 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.

[25] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, vol. 2, pp. 83–97, 1955.

[26] 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.