Robot-Centric Activity Prediction from First-Person Videos: What Will They Do to Me?

M. S. Ryoo[1], Thomas J. Fuchs[1], Lu Xia[2,3], J. K. Aggarwal[3], Larry Matthies[1]
[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
[2]Amazon.com, Inc., Seattle, WA
[3]Department of ECE, the University of Texas at Austin, Austin, TX
mryoo@jpl.nasa.gov

ABSTRACT

In this paper, we present a core technology to enable robot recognition of human activities during human-robot interactions. In particular, we propose a methodology for *early recognition* of activities from robot-centric videos (i.e., first-person videos) obtained from a robot's viewpoint during its interaction with humans. Early recognition, which is also known as *activity prediction*, is an ability to infer an ongoing activity at its early stage. We present an algorithm to recognize human activities targeting the camera from streaming videos, enabling the robot to predict intended activities of the interacting person as early as possible and take fast reactions to such activities (e.g., avoiding harmful events targeting itself before they actually occur). We introduce the novel concept of 'onset' that efficiently summarizes pre-activity observations, and design a recognition approach to consider event history in addition to visual features from first-person videos. We propose to represent an onset using a cascade histogram of time series gradients, and we describe a novel algorithmic setup to take advantage of such onset for early recognition of activities. The experimental results clearly illustrate that the proposed concept of onset enables better/earlier recognition of human activities from first-person videos collected with a robot.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*video analysis*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*motion*; I.2.9 [**Artificial Intelligence**]: Robotics—*sensors*

Keywords

Activity recognition; first-person videos; human-robot interaction

1. INTRODUCTION

First-person activity recognition is a research area studying automated recognition of human activities from videos with the actor's own viewpoint. Its main difference to the conventional 3rd-person activity recognition is that the observer wearing the camera (e.g., a robot) himself/herself is involved in the ongoing activity, making its perception to become egocentric videos. This makes this research area very germane to robotics (i.e., any visual observation from a viewpoint of a robot becomes a first-person video), understanding of natural human-robot interactions in particular. First-person activity recognition is essential to provide a robot activity-level situation awareness 'during' social and physical human-robot interactions (e.g., recognizing a hostile interaction that a human is punching the robot and the robot is collapsing as a consequence). Such concept is very different from conventional robot recognition of human gesture-level commands (e.g., [18, 2]) or task-level actions (e.g., [9]) from stationary cameras without any direct physical interaction between humans and the robot. In our scenarios, the camera moves dynamically as the robot gets involved in interactions (e.g., Figure 4) and videos visually display very different characteristics compared to conventional scenarios.

An ability particularly important and necessary for first-person recognition systems is the ability to infer humans' intended activities at their early stage. For instance, public service robots and surveillance/military robots must protect themselves from any harmful events by inferring the beginning of dangerous activities like an 'assault'. Similarly, a wearable system must recognize ongoing events around the human as early as possible to provide appropriate service for human tasks and to alarm accidents like 'a car running into the person'. Natural human-robot interaction also becomes possible by making robots to provide early reaction to humans' actions. This is not just about real-time implementations of activity recognition, but more about recognition of activities from observations only containing the beginning part of the activity. The objective is to detect an ongoing activity in the middle of the activity execution, before it is completed.

This problem, recognition of an activity before fully observing its execution, is called 'early recognition' or 'activity prediction' [14]. However, even though there are recent works on early recognition, (1) it has never been studied for first-person videos and (2) research on an early recognition approach that simultaneously considers pre-activity observations as well as features from the ongoing activity has been limited. In real-world first-person recognition scenarios, the system is required to continuously process long video inputs containing a sequence of multiple activities. As a consequence, it is important for the system to analyze not only the video segment corresponding to the ongoing activity but also other activity history or signals observed 'prior' to the beginning of the activity. In this paper, we call such signals observed before the activity as an *onset* of the activity.

This paper newly introduces the concept of onset, and presents an early recognition approach to take advantage of them for the robot recognition. We formulate the early recognition (i.e., pre-

HRI'15, March 2–5, 2015, Portland, Oregon, USA.
Copyright © 2015 ACM 978-1-4503-2883-8/15/03 ...$15.00.
http://dx.doi.org/10.1145/2696454.2696462.

diction) problem to consider activity history and human intention together with ongoing observation of the activity, and discuss how our *onset signatures* enable abstraction of such pre-activity observations for better recognition of activities. We define an onset activity as short and subtle human motion (e.g., waving and reaching) observable before main activities (e.g., shaking hands and throwing an object), and attempt to capture/model onset patterns displayed prior to each main activity. More specifically, we compute a collection of weak classifier responses (each corresponding to a particular onset activity) over time and construct cascade histograms of their time series gradients as our representation summarizing pre-activity observations: onset signatures. Our method is particularly designed to capture loose stochastic correlations between the onset and the target activities (e.g., reaching an object may or may not occur before throwing but they are correlated) and also consider absence of a certain onset (e.g., absence of waving before punching) for better recognition. An efficient (linear time complexity) algorithm is designed to take advantage of our onset signatures to perform better early recognition from continuous videos, while minimizing the amount of computations.

We formulate the early recognition problem in Section 2. We present the concept of onset and our recognition approach to utilize them in Section 3. Experimental results are discussed in Section 4, and Section 5 concludes the paper.

1.1 Related work

The research area of first-person activity recognition is gaining an increasing amount of attention recently. There are several works on recognition of ego-actions of the person (i.e., actions of the person wearing a camera such as skiing) [7, 4], object-oriented analysis of humans using objects (e.g., a towel) [12, 13], or analysis based on face and gaze [5]. However, only very few works considered recognition of interaction-level activities where multiple humans (or robots) physically interact each other [16]. Furthermore, no previous work attempted early recognition from first-person videos.

The problem of early recognition (i.e., activity prediction) was introduced and formulated with modern spatio-temporal features in [14], but it was limited to 3rd-person videos and did not consider pre-activity observations. There also have been works considering past activity history for predicting future states/locations using state-models [9] and/or trajectories [8, 20] from 3rd-person videos. However, even though these approaches are appropriate for predicting future steps of the activities composed of clear states, they are unsuitable for directly handling dynamic first-person videos whose analysis requires various types of spatio-temporal video features [11, 3, 17] that display highly sparse and noisy characteristics. In order to enable accurate early recognition for interaction-level first-person activities, simultaneous consideration of pre-activity observations (i.e., an onset) and ongoing activity observations is needed.

To our knowledge, this paper is the first paper is discuss 'early recognition' problem for *first-person* videos from robots. As pointed out above, previous activity prediction works were designed for videos obtained from a static camera [14, 8] or those from a robot simply standing without any ego-motion or interaction [9]. This paper newly discusses the problem of activity prediction with first-person videos displaying robot ego-motion (e.g., rotating) as well as camera motion (e.g., shaking) caused by physical human-robot interactions. We also believe it is the first paper to explicitly consider *pre-activity observations* (i.e., frames 'before' the starting time of the activity) for early recognition, which was not attempted in [14, 6].

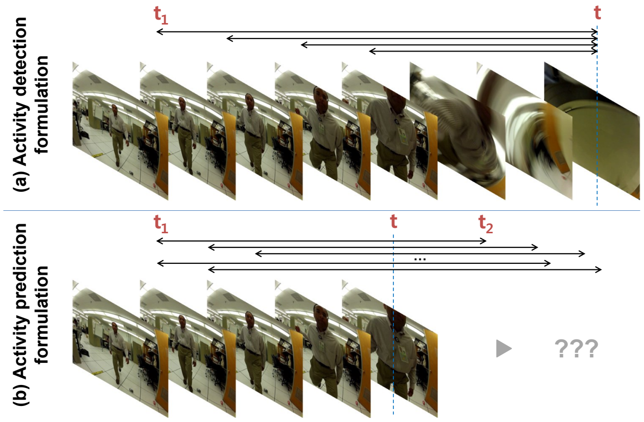

(a) Activity detection formulation
t1
t
(b) Activity prediction formulation
t1
t
t2
...
???

Figure 1: Time intervals that need to be considered in our activity detection/prediction formulations. t indicates the observation time (i.e., the current frame) and t_1 and t_2 specify starting and ending times of time intervals internally considered.

2. PROBLEM FORMULATION

2.1 Basic activity detection framework

Human activity detection is the problem of finding starting time and ending time (i.e., a time interval) of each occurring activity from a video input. Given a continuous video stream from frame 0 to t (we denote this as $V[0, t]$ or simply as V), for each activity class C, the system is required to obtain time intervals $[t_1, t_2]$ that it believes to correspond to C. In general, this is solved by evaluating all possible time intervals (i.e., $[t_1, t_2]$ where $0 \leq t_1, t_2 \leq t$) or estimating intervals providing local maximum probability values: $P(C^{[t_1,t_2]}|V)$. Here, $C^{[t_1,t_2]}$ denotes the event of the activity C starting exactly at time t_1 and ending exactly at t_2.

Assuming that the robot is only interested in the 'activity that just happened', this can be further simplified as:

$$\begin{aligned} P(C^t \mid V) &= \sum_{t_1} P(C^{[t_1,t]} \mid V) \\ &= \frac{\sum_{t_1} P(V[t_1,t] \mid C^{[t_1,t]}) P(C^{[t_1,t]})}{\sum_{C,t_1} P(V[t_1,t] \mid C^{[t_1,t]}) P(C^{[t_1,t]})} \end{aligned} \tag{1}$$

where $P(C^t|V)$ represents the probability of the activity 'ending' at time t regardless of its starting time, and we use a basic probability marginalization with possible starting times t_1. We call this problem more specifically as 'after-the-fact detection', since the system only focuses on the activity which is already finished at t (Figure 1 (a)). Recognition can be performed by computing the above probability at every time step as the video continues.

Instead of directly using the video stream V, visual features abstracting V are often extracted and used for the recognition [1]. These include sparse local motion descriptors [11] capturing salient local movements observed in videos, global motion descriptors [16] representing camera movements, and semantic descriptors like per-frame human body poses [19]. By modeling the distributions of such features corresponding to the activity (i.e., $P(V|C)$), the recognition can be performed following Equation 1: this is a binary classification of deciding whether the activity C is occurring or not at each frame t. In our case, the concept of bag-of-visual-words were used as our feature representation, modeling each distribution as a histogram similar to [16]. Also notice that each visual feature (e.g., 3-D XYT volume patch in the case of [11]) is extracted with a particular time stamp, and they stay as is once extracted.

2.2 Activity prediction formulation

The 'activity prediction' problem is the problem of recognizing ongoing activities at their early stage. In contrast to the above after-the-fact detection problem, recognition must be made in the middle of the activity before it is fully executed. The system must consider the possibility that the activity is 'ongoing' at frame t, thus considering time intervals where $t_1 \leq t \leq t_2$ (Figure 1 (b)). In addition, the system is required to explicitly consider multiple progress levels d of the activity C:

$$P(C^t|V) = \frac{\sum_d \sum_{[t_1,t_2]} P(V[t_1,t] \mid C^{[t_1,t_2]}, d) P(C^{[t_1,t_2]}, d)}{\sum_{C,d,[t_1,t_2]} P(V[t_1,t] \mid C^{[t_1,t_2]}, d) P(C^{[t_1,t_2]}, d)} \quad (2)$$

where t_2 is a future frame and observation corresponding to $V[t+1, t_2]$ is not available. The variable d is a conceptual progress status of the activity (i.e., up to which point the activity has progressed so far?) having a value between 0 and 1. Assuming that each activity progresses linearly when occurring, the following equation holds: $t = t_1 + d \cdot (t_2 - t_1)$.

We also call this as early 'detection' problem, which extends the early 'classification' problem (i.e., early categorization of segmented videos) introduced in [14].

Early detection of human activities with context: Even though the above formulation enables early detection of activities, it is often insufficient for continuous video scenarios. It only utilizes the video segment corresponding to the time interval alone (i.e., $V[t_1, t]$) to make the decision, while ignoring all the other previous video observations (i.e., $V[0, t_1 - 1]$). In continuous videos, activities occur in a sequence and they are correlated. Furthermore, the interacting person usually has his/her own intention, such as 'harming' the camera or 'avoiding' the robot. Figure 2 illustrates a graphical model describing such activity-activity relations and intention-activity relations.

Thus, the early detection problem can be formulated as:

$$\begin{aligned} P(C^t \mid V) = & \sum_d \sum_{[t_1,t_2]} P(C^{[t_1,t_2]}, d \mid V) \\ \propto & \sum_d \sum_{[t_1,t_2]} \sum_{(\mathbf{A},I)} P(C^{[t_1,t_2]}, d, \mathbf{A}, I, V) \end{aligned} \quad (3)$$

where $t_1 \leq t \leq t_2$, $\mathbf{A}$ is a set of all previous activities, I is the intention of the interacting person, and $P(C^{[t_1,t_2]}, \mathbf{A}, I, V)$ is the joint probability. In our approach, we use the function $F(\cdot)$ to approximate this joint probability term, while designing it to consider activity-activity relations (i.e., $\mathbf{A} = \{A_1, \cdots, A_{|\mathbf{A}|}\}$ and C) as well as intention-activity relations displayed in Figure 2.

The key issues for the early detection are (i) designing the robust joint probability function F, (ii) learning the model parameters in F from training videos, and (iii) making an inference given a new video observation V. This inference must be made at its every time frame t while considering possible intervals $[t_1, t_2]$. We emphasize that t is smaller than t_2 in the case of an ongoing activity (i.e., it is in the middle of execution), and F must be designed to consider such characteristic. We discuss this more in Section 3.3.

Challenges: The main technical challenge is that the above computations must be performed in real-time, making the detection as early as possible. Particularly in robot perception, it is very contradictory to say that "even though the approach is able to perform early recognition, its processing time will take multiple seconds/minutes". This implies that (1) we need to apply the recognition algorithm almost every frame (i.e., it should not wait) and that (2) the algorithm still must perform in real-time or faster. This makes a standard way of modeling/training the function F and making an multi-iteration inference using a latent SVM formulation similar to [10] or MCMC-based searching difficult.

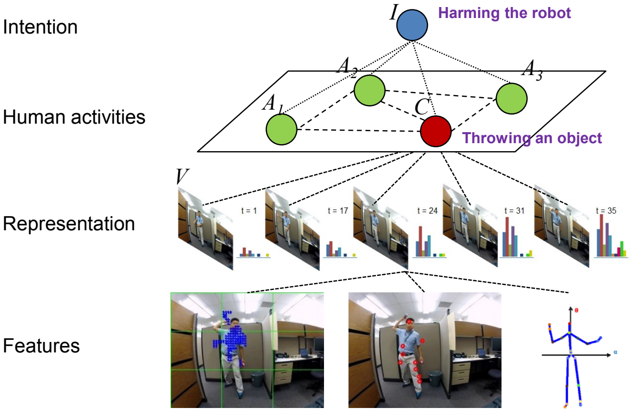


Figure 2: Graphical model representation of the scenario where an interacting human performs a sequence of activities under a specific intention. The robot is required to consider activity-activity relations, capturing pre-activity observations.

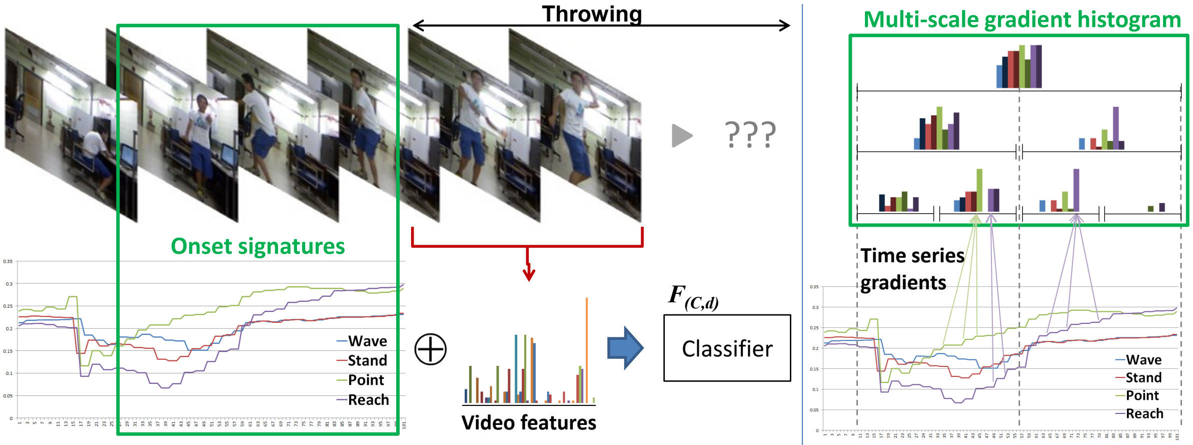
3. ACTIVITY PREDICTION USING ONSET

In order to enable early detection while addressing the above mentioned challenges, we introduce the new concept of 'onset activities' and 'onset signatures' together with our recognition approach to take advantage of them. The idea is to learn weak detectors for subtle short-term activities (i.e., onset activities) which are closely or loosely related to the occurrence of activities-of-interest, and make the recognition system to capture activity-activity relations (i.e., $\mathbf{A}$) using such onset information. Our approach learns onset patterns leading to the occurrence of each target activity while explicitly considering stochastic nature of onsets, and performs its early recognition by analyzing onset distributions observed before the activity. Figure 3 (a) illustrates its overall concept.

3.1 Onset activities

We define *onset activities* as subtle activities which (1) occur within a short time duration and (2) do not physically influence the observer (i.e., a robot or a wearable camera), but (3) serve as a direct/indirect cue to infer their following activities. 'Standing', 'pointing', and 'picking up an object' are typical examples of onset activities. These activities themselves do not have strong meaning and they do not influence the robot/camera directly, but they can serve as indicators describing 'what activity is likely to follow next'. An example will be the activity of 'picking up an object' serving as the onset for 'throwing an object'. Another example will be 'waving' before 'hand shaking'.

Typically, because of the subtle nature of onset activities, their recognition becomes difficult and unreliable. The activities usually contain a small amount of human motion (e.g., only a subtle arm/finger gesture is visible when 'pointing'). This makes the detectors for onset activities to become weak classifiers, and prevents the system from directly using the onset recognition results. For instance, average precisions (AP) of our onset activity detection were 0.1 to 0.2 in our dataset. Furthermore, onset activities often have very stochastic nature, implying that onsets are not always observable before activities (e.g., the person may have the ball beforehand and throw it without the picking up action). Thus, an approach to best utilize these weak detectors is required.



(a) Early activity recognition framework with onset

(b) Onset signature representation

Figure 3: (a) Illustration of the overall concept of our early activity recognition pipeline using onset. The recognition approach is designed so that it considers both the pre-activity observations (onset signatures) and video features from ongoing activities. (b) Illustration of our onset signature representation, based on the multi-scale cascade histogram of time series gradients.

3.2 Onset signatures

We define *onset signatures* as a set of time series describing onset activity matching results. That is, given a continuous (streaming) video, we measure the similarity between each possible video segment and the onset activity, and record how the similarity is changing over time. The objective is to use these time series as features suggesting future activities. Each onset signature $G^k(t)$ of kth onset activity is more specifically computed as:

$$G^k(t) = \max_r (1 - D^k([t-r, t])) \tag{4}$$

where r is the model duration of the activity, and $D^k([t-r, t])$ is the distance between the model of the kth onset activity and the video observation segment $V[t-r, t]$. We use a basic template matching of bag-of-words representations (obtained from a set of training videos S^k) as our D^k:

$$D^k([t_1, t_2]) = \sum_i (m_i^k - v_i[t_1, t_2]))^2 \tag{5}$$

where v_i is the ith feature value and m_i^k is its mean model value: $m_i^k = \sum_{V^j \in S^k} v_i^j / |S^k|$. In our implementation, we took advantage of the same video features (with bag-of-visual-words representation) described in Section 2.1: [11, 3, 19, 16].

The matching is performed for all t and possible r values, providing us the final $G^k(t)$. The resulting $G^k(t)$ forms a time series, describing how our onset activity detector is responding to the ongoing video observation. We collect $G^k(t)$ from all onset activities and use them as our *onset signature*.

The template matching process takes time complexity of $O(n \cdot R)$ per frame where n is the feature dimension and R is the number of temporal window sizes we consider ($R < 5$ in our case). Furthermore, the computation of onset signature per activity is independent to each other, making its parallelization possible.

Histogram representation of onset signatures: We design a histogram representation of onset signatures (Figure 3 (b)). The idea is to make the system efficiently summarize the previous onset occurrence information from its time series, so that it can use it to infer ongoing/future activities.

Typical representations of onset signatures are mean and maximum values of a fixed time window (e.g., frames between the current frame and 50 frames before that). However, this is often insufficient due to noisy and weak nature of onset matching (notice that onset recognition relying on peak matching values give us $\sim$0.1 AP), and deeper analysis of time series is necessary. Thus, we construct cascade histograms of time series gradients to represent onset signatures.

Let $||$ denote the concatenation operation of two vectors, $[a_1, \cdots, a_n] \; || \; [b_1, \cdots, b_n] = [a_1, \cdots, a_n, b_1, \cdots, b_n]$. Then, the histogram representation of onset signature H at time t is defined as: $H(t) = H_1(t-u, t) \; || \; H_2(t-u, t) \; || \; \cdots \; || \; H_{|\mathbf{A}|}(t-u, t)$, where $H_k(t-u, t)$ is the histogram for the kth onset activity computed based on the time interval $[t-u, t]$ with duration u (e.g., 50). H_k is defined more specifically as:

$$\begin{aligned} H_k(t_1, t_2) = & H_k\left(t_1, \frac{t_1 + t_2}{2}\right) \; || \; H_k\left(\frac{t_1 + t_2}{2}, t_2\right) \\ & || \; [h_k^+(t_1, t_2), h_k^-(t_1, t_2)] \end{aligned} \tag{6}$$

where

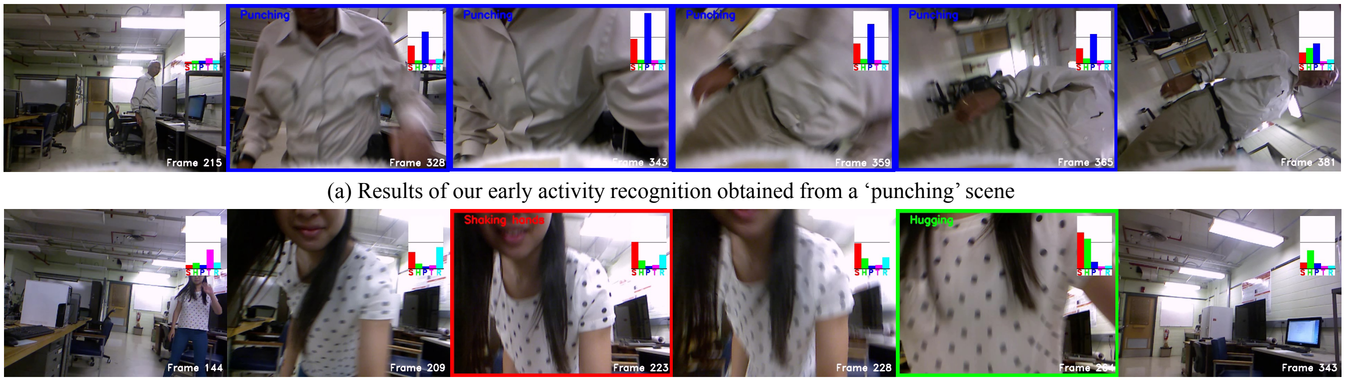
$$\begin{aligned} h_k^+(t_1, t_2) &= \left| \{ t_1 \le t \le t_2 \mid G^k(t) - G^k(t-s) > 0 \} \right|, \\ h_k^-(t_1, t_2) &= \left| \{ t_1 \le t \le t_2 \mid G^k(t) - G^k(t-s) \le 0 \} \right|. \end{aligned} \tag{7}$$

Here, s is the step size of gradient computation, and we perform this histogram construction for multiple s scales and concatenate the results.

The above recursive equation hierarchically performs temporal segmentation of the time series (i.e., our onset signatures) into multiple parts, and obtains a histogram of time series gradients corresponding to each of them. That is, our hierarchical histogram is constructed by applying our recursive function until it reaches the level l. In our experiments, $l = 3$ gave us good results.

The final feature vector representation of the onset signature is constructed as follows, by attaching mean and max values to the histogram:

$$\begin{aligned} x(t) = H(t) \; || & \left[\sum_{t'=t-d}^{t} \frac{G^1(t')}{u}, \cdots, \sum_{t'=t-d}^{t} \frac{G^n(t')}{u} \right] \\ & || \; \left[\max(G^1(t')), \cdots, \max(G^n(t')) \right]. \end{aligned} \tag{8}$$



(a) Results of our early activity recognition obtained from a ‘punching’ scene

(b) Results of our early activity recognition obtained from a ‘hand shaking’ - ‘hugging’ scene

Figure 4: Example result image sequences of our early activity recognition. A ‘punching’ activity (blue boxes), a ‘hand shaking’ activity (red box), and a ‘hugging’ activity (green box) are detected. We are able to observe that the camera displays ego-motion due to human-robot interactions, such as it collapsing due to ‘punching’ and it shaking during ‘hand shaking’ interaction.

3.3 Early detection using onset signature

Based on its video observation V and computed onset signatures x, our approach performs early detection of an activity by using a set of binary classifiers. More specifically, we formulate the detector at time t as:

$$\begin{aligned} P(C^t \mid V) &\propto \sum_{d} \sum_{[t_1,t_2]} \sum_{(\mathbf{A},I)} P(C^{[t_1,t_2]}, d, \mathbf{A}, I, V) \\ &\propto \sum_{d} \sum_{[t_1,t_2]} \sum_{(\mathbf{A},I)} P(\mathbf{A}, V | C^{[t_1,t_2]}, d) \cdot P(C^{[t_1,t_2]}, d | I) \\ &\approx \max_{d} \max_{[t_1,t_2]} \sum_{I} F_{(C,d)}(x(t), V[t_1,t]) \cdot L_C([t_1,t_2], I) \end{aligned} \tag{9}$$

where we factor the joint probability into two terms using conditional independence and uniform prior assumptions. The functions F and L_C are used to estimate the two terms while explicitly reflecting $t_1 \leq t \leq t_2$. We use support vector machine (SVM)-based probability estimation in our implementation to approximate the terms, which we describe more below.

We trained one binary classifier for each $F_{(C,d)}$ and made it to estimate the probability scores. To support Equation 9, this was done for each pair of activity C and possible progress level d. A concatenation of the vector describing video features inside the interval (i.e., $V[t_1,t]$) and the vector representation of our onset signature (i.e., $x(t)$) serves as an input to these classifiers, and each of the learned classifier $F_{(C,d)}$ measures the probability of the activity C ongoing at the time t (with progress level d). The training of the classifier is performed by providing positive and negative samples of V and x together with their ground truth labels y. The function L_C is trained similarly, by providing I and $[t_1,t_2]$ as inputs.

The idea is to abstract previous activity occurrences (i.e., $\mathbf{A}$) using our onset signature x, instead of making the system to enumerate through all possible combinations. That is, we directly used our onset signatures as input features of the function F, thereby enabling efficient computations. Based on the training samples, the classifier will learn to focus on particular onset signature while ignoring irrelevant onset activities.

Overall computations required for activity recognition is $O(n \cdot |d| \cdot R)$ at each time step if a binary classifier with a linear complexity is used (e.g., SVM), where n is the feature dimension, $|d|$ is the number of possible activity progress levels (we used 10 levels in our experiments), and R is the number of activity durations we consider (this influences possible starting points of the time interval, t_1). Our approach is able to cope with any types of binary classifiers in principle (by making them predict either 0 or 1), and does it more reliably with classifiers estimating probability.

4. EXPERIMENTS

4.1 Dataset

We constructed a new dataset composed of continuous human activity videos taken from a robot’s first-person viewpoint. It is an extension of the previous humanoid-based first-person video dataset [16] whose videos mostly contain a single activity; our new videos contain a sequence of 2∼6 activities (onset activities and interactions). The motivation is that the community has been lacking a public dataset for ‘early recognition’ (i.e., activity prediction): To our knowledge, the public dataset most commonly used for activity prediction is UT-Interaction [15]. However, even if we set aside that UT-Interaction is not a first-person video dataset, it has a major limitation: its videos contain activities executed in an random order without any context (e.g., punching and then shaking hands). This is very unnatural, since the actors are following a fixed script without any intention on their own (unlike our new dataset).

Our camera was mounted on top of a humanoid similar to [16], and we asked human subjects to perform a series of activities with three different types of intentions: friendly, hostile (i.e., harming), and avoiding. We labeled 9 types of human-robot interactions performed by the actors: 4 types of onset activities and 5 types of activities-of-interest. ‘Pointing the observer (i.e., the robot camera)’, ‘reaching an object’, ‘standing up’, and ‘waving to the observer’ are the 4 onset activities. ‘Handshaking with the observer’, ‘hugging the observer’, ‘punching the observer’, ‘throwing an object at the observer’, and ‘running away from the observer’ are the 5 interaction-level activities in our dataset. The humanoid moves as it is involved in the interaction (e.g., the camera jolting as an impact of throwing, and the camera collapsing as a result of punching). Furthermore, our experiments were designed to capture translation and rotation movements of our robot, and thus these robot ego-motion is present in our first-person videos. Figure 4 shows example frames of our videos.

Videos with the resolution of 640*480 with 30fps were used. Also, 320*240 depth videos were collected in addition to the conventional RGB videos, so that the system obtains Kinect-based pos-

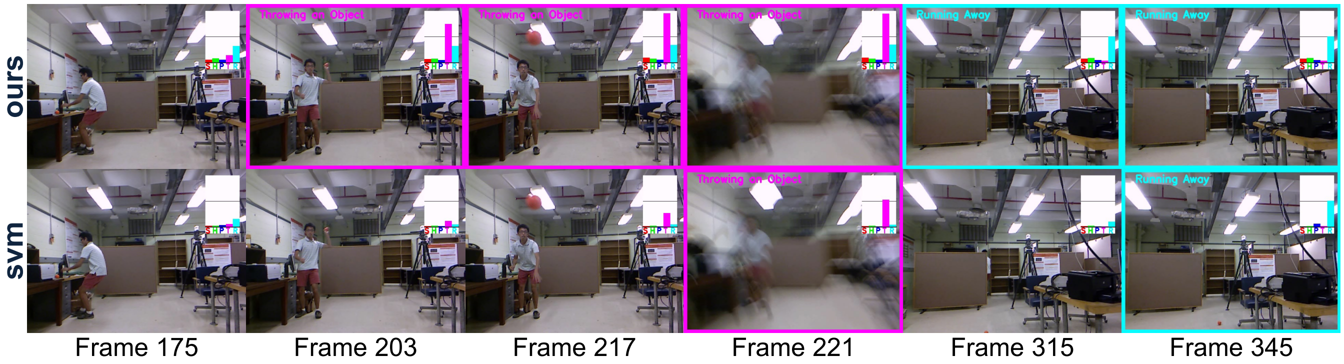


Figure 5: Example result image sequences comparing our early detector (top) with the previous SVM detector using state-of-the-art features (bottom). A 'throwing' activity (magenta boxes) and a 'running away' activity (cyan boxes) are detected. Notice that the previous method detected the 'throwing' only *after* the ball actually hit the camera, while ours detected it as soon as the person raised the hand. That is, our approach detects activities at their much earlier stage compared to the previous detector.

ture estimation results when available. The dataset consists of 8 sets, where each set contains continuous videos of human activities being performed by the same subject. It contains a total of 61 continuous videos with ∼180 executions of human activities.

4.2 Implementation

We extracted multiple types of state-of-the-arts visual features from first-person videos, including global motion descriptors [16], local motion descriptors [11, 3], and human posture descriptors [19]. Once features are extracted from raw videos, we clustered these features to obtain standard bag-of-visual-words feature representation while using integral histograms for more efficient computations. Our approach and multiple baseline approaches including the state-of-the-art early recognition approach [14] were trained/tested/compared using our dataset.

We implemented (1) our approach taking advantage of onset signatures as well as (2) its simplified version designed to only use peak onset activity responses (instead of full onset signatures). In addition, we implemented (3) an extended version of previous state-of-the-art early recognition approach [14] originally designed for the 3rd-person videos, and (4) made it to also take advantage of our onset signatures. Furthermore, we implemented several baseline activity detection approaches including (5) a sliding window detector with Bayesian classifiers assuming a Gaussian distribution (i.e., a after-the-fact detection approach), and (6) a state-of-the-art activity detector using SVM with a non-linear kernel (i.e., RBF). All of the above approaches took advantage of the same features vector (i.e., the concatenation of four feature types [11, 3, 19, 16]), which outperformed those using single feature type. We also tested (7) the approach detecting activities solely based on onset signatures (i.e., context-only).

All these approaches run faster than real-time on a standard desktop PC with our unoptimized C++ implementation (0.0036 sec per frame), except for the adopted feature extraction part.

4.3 Evaluation

We use leave-one-set-out cross validation (i.e., 8-fold cross validation) for continuous 'detection' tasks. Ground truth labels of activity occurrences in videos for both the onset activities and the interaction activities were provided, so that we can take advantage of them for the training (i.e., a supervised learning setting) and testing. At each round, for the testing, our approach computes the probability $P(C^t \mid V)$ (which also can be viewed as a confidence score) at every frame t. Treating its peak values as detections (while discarding overlapping intervals), we computed precision-recall curves (PR-curves) by changing the detection threshold. Detected time intervals that overlap more than 50% with the ground truth activity intervals were considered as true positives. Average precision (AP) is also obtained from the curve by measuring the area under the PR-curve, and mean AP is computed by averaging APs of all activity classes.

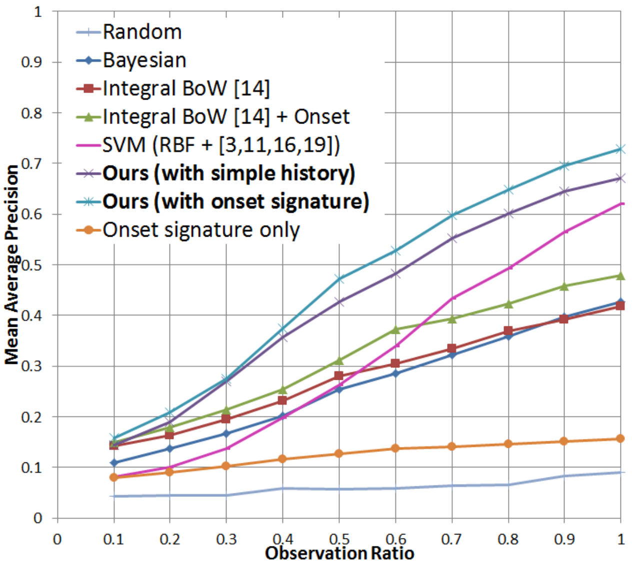


Figure 6: A figure comparing performances of our early detection approaches with previous works and baselines. Mean AP, which is an area under a precision-recall curve, is measured per observation ratio. A higher graph suggests better performance; a higher graph indicates that it 'recognizes activities more accurately given the same amount of observation' and that 'it is able to recognize activities earlier than the others if the same accuracy is assumed'. It clearly shows superiority of ours.

In addition, in order to measure the early detection ability of our approach, we tested our approaches and baselines with multiple different observation ratio settings similar to [14]. More specifically, activity observation ratio was set from 0.1 to 1.0, and mean AP was measured per observation ratio. An observation ratio specifies the progress level of the activity execution. For example, observation

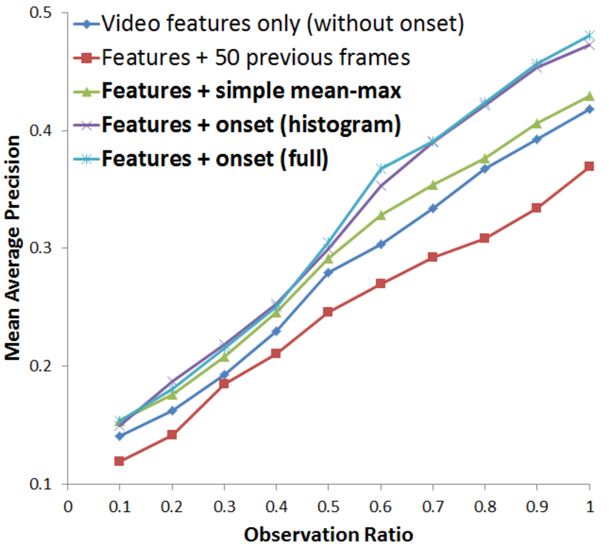


Figure 7: We compared approaches using different onset representations to illustrate the advantages of our proposed cascade histogram-based onset signature representation, while using an approach without randomness [14] as a baseline.

ratio of 0.2 implies that the system was asked to make the detection decision after observing the initial 20% of the activity (i.e., very early detection), and observation ratio of 1.0 implies that it is a standard after-the-fact detection. Let d be the observation ratio and $[g_1, g_2]$ be a ground truth time interval of the activity. For each experiment with an observation ratio d, we removed all video features extracted from the time interval $[g_1 + d \cdot (g_2 - g_1), g_2]$ (i.e., those observed after d). In addition, only the detection found before the observation ratio were consider as true positives.

4.4 Results

Figure 6 shows the mean AP values of activity detectors measured with various observation ratio settings. We are able to observe that our activity prediction formulation of using onset activities and their signatures is benefiting the system greatly. Mean APs of our approach (with onset) were constantly higher by 0.1∼0.2 compared to the baseline SVM classifier using state-of-the-art features, achieving the same AP much earlier. For instance, in order to obtain the mean AP of 0.5, our early detector with onset signatures requires 55% observation while the SVM requires more than 80%. This gap can also be observed for integral bag-of-words with and without onset. Figure 5 shows example images of these two detection results, confirming the superiority of our proposed approach.

Figure 8 illustrates PR curves of the approaches. Early recognition approaches with our onset signatures particularly performed well on the activity of 'throwing an object', since it very often had a clear onset activity: 'reaching the object'. Our approach also performed well for 'hugging' and 'shaking' (relying on the existence of the onset 'waving' and the absence of 'reaching' or 'pointing'), and detected 'punching' earlier than those not using onset.

We also conducted an additional experiment to investigate advantages of our cascade histogram-based onset signature representation. We compared the performance of our onset representation with various other onset representations, including (i) the approach adding video features obtained 1∼50 frames prior to the activity in addition to those from the activity's actual video segment, (ii) the approach using a simple onset representation of mean and max values, (iii) the approach only using our histogram-based onset representation, and (iv) our final onset representation composed of histogram + mean and max. Figure 7 illustrates the result. It clearly shows that our onset signature representation effectively captures previous video information. Particularly, we are able to observe that simply adding 50 frames prior to the time interval is only confusing the system. Integral BoW was used as the base classifier in this experiment, since it does not contain randomness.

Table 1: A table comparing performances of the proposed approach with state-of-the-arts in terms of mean average precision (AP). Our proposed approach outperformed previous works particularly when making early recognition.

Method	50% observation	Full observation
Ours	**0.473**	**0.729**
Ryoo et al. [16]	0.360	0.717
Integral BoW [14]	0.280	0.418
SVM [3,11,16,19]	0.263	0.620
Bayesian	0.254	0.427
Onset only	0.127	0.156
Random	0.056	0.089

Finally, we explicitly compared recognition performances of our proposed early recognition approach with previous state-of-the-art approaches on our video dataset. Table 1 shows the results. Not only the final activity detection accuracies but also the early detection accuracies (i.e., observation ratio 50%) were compared.

5. CONCLUSION

This paper presented a methodology for early recognition of human activities. Early recognition ability is very essential for first-person vision systems which are required to function in real-world environments in real-time while constantly interacting with others. This makes the proposed technology very necessary for robust human-robot interactions, and this paper investigates such concepts for the first time. We formulated the early recognition problem to consider pre-activity observations, and presented an efficient new approach that uses the concept of 'onset'. Experimental results confirmed that our formulation enables superior early recognition performance to previous conventional approaches, and that our histogram-based representation of onset signatures benefits early recognition by capturing pre-activity observations.

Acknowledgment: The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0016.

6. REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43:16:1–16:43, April 2011.

[2] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE T PAMI*, 31(9):1685–1699, 2009.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE Workshop on VS-PETS*, 2005.

[4] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011.

[5] A. Fathi, J. Hodgins, and J. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012.

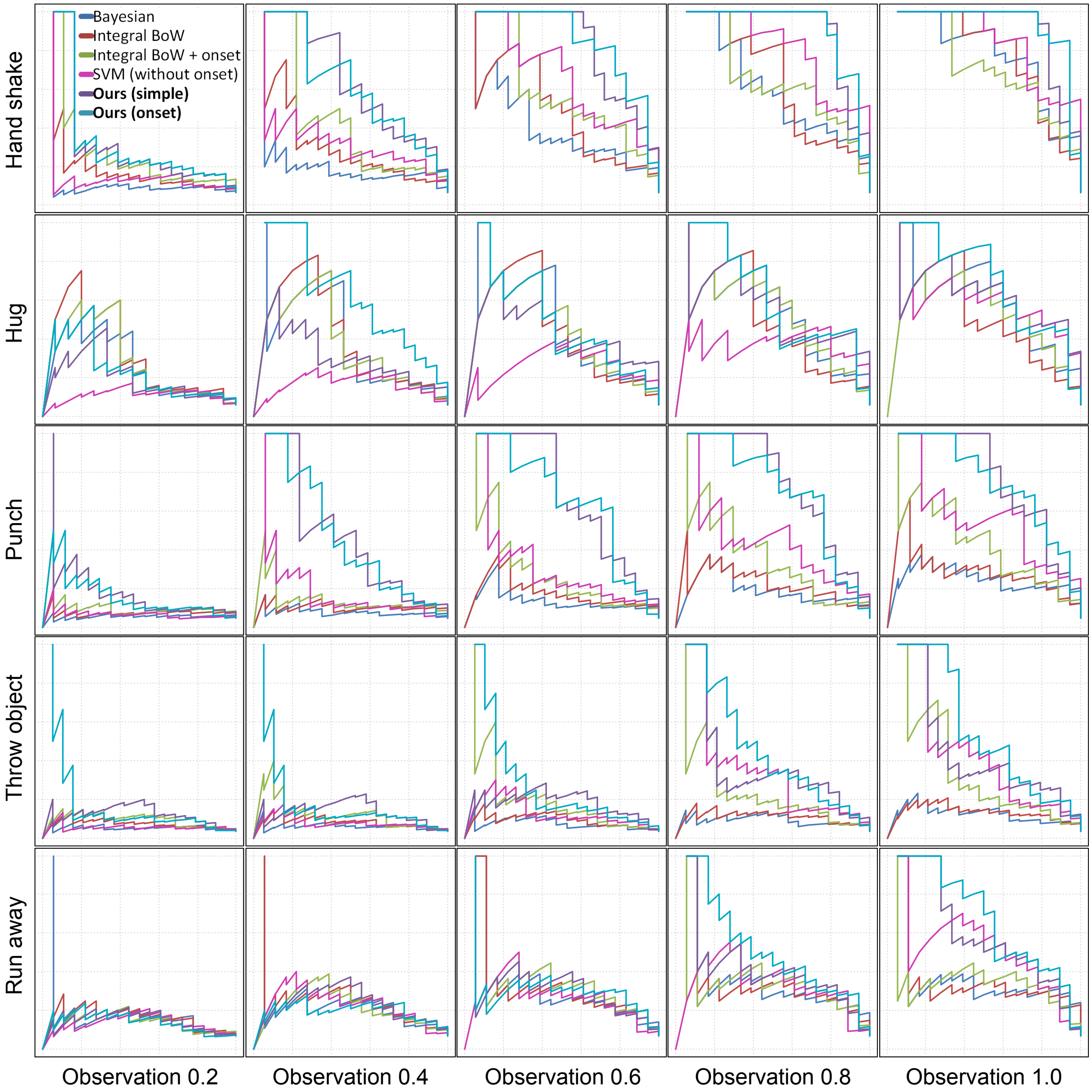


Figure 8: Precision-recall curves of each activity per observation ratio setting. X axis [0∼1.0] of all graphs is 'recall' and Y axis [0∼1.0] is 'precision'. Our approach with onset showed the best performance in all cases. Using our onset signature (light blue) particularly showed a huge performance increase over SVM (pink).

[6] M. Hoai and F. D. la Torre. Max-margin early event detectors. In *CVPR*, 2012.

[7] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011.

[8] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012.

[9] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *RSS*, 2013.

[10] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012.

[11] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.

[12] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.

[13] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[14] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011.

[15] M. S. Ryoo and J. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009.

[16] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013.

[17] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[18] S. Waldherr, S. Thrun, and R. Romero. A gesture-based interface for human-robot interaction. *Autonomous Robots*, 9(2):151–173, 2000.

[19] L. Xia, C.-C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *CVPRW*, 2012.

[20] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring "dark matter" and "dark energy" from videos. In *ICCV*, 2013.