Prediction of Human Activity by Discovering Temporal Sequence Patterns

Kang Li, *Student Member, IEEE* and Yun Fu, *Senior Member, IEEE*

Abstract—Early prediction of ongoing human activity has become more valuable in a large variety of time-critical applications. To build an effective representation for prediction, human activities can be characterized by a complex temporal composition of constituent simple actions and interacting objects. Different from early detection on short-duration simple actions, we propose a novel framework for *long*-duration complex activity prediction by discovering three key aspects of activity: **Causality**, **Context-cue**, and **Predictability**. The major contributions of our work include: (1) a general framework is proposed to systematically address the problem of complex activity prediction by mining temporal sequence patterns; (2) probabilistic suffix tree (PST) is introduced to model causal relationships between constituent actions, where both large and small order Markov dependencies between action units are captured; (3) the context-cue, especially interactive objects information, is modeled through sequential pattern mining (SPM), where a series of action and object co-occurrence are encoded as a complex symbolic sequence; (4) we also present a predictive accumulative function (PAF) to depict the predictability of each kind of activity. The effectiveness of our approach is evaluated on two experimental scenarios with two data sets for each: action-only prediction and context-aware prediction. Our method achieves superior performance for predicting global activity classes and local action units.

Index Terms—Activity prediction, causality, context-cue, predictability

1 Introduction

In recent years, research shows that modeling temporal structure is a basic methodology for recognition of complex human activity [20], [7], [41]. These studies extend the types of human activity that can be understood by machine vision systems. Advances in this field made an important application become real: *predicting activities or imminent events from observed actions or events in the video*. Many intelligence systems can benefit from activity prediction. For instance, in the sports video analysis, the capability of predicting the progress or results of a sports game is highly desirable. In public areas, we want to equip a surveillance system that can raise an alarm in advance of any potential dangerous activity happens. In a smart room, people's intention of activity can be predicted by a user-friendly sensor—camera, so that the system will adaptively provide services, even help if necessary.

Though activity prediction is a very interesting and important problem, it is quite a new topic for the domain of computer vision. To the best of our knowledge, the work in [45] is the only one that explicitly focused on this problem. They identified activity prediction with early detection of short-duration single action, such as "hugging", "pushing". This assumption limits the types of activities that can be predicted as well as how early the prediction can be made. We believe that activity prediction is more desirable and valuable if it focuses on long-duration complex activities, such as "making a sandwich". The early detection problem can be solved in the classic recognition paradigm by predicting directly on low-level feature representations. Our approach aims to solve the long duration prediction problem with a completely different framework, where semantic-level understanding and reasoning are our focus.

Specifically, in this paper, we propose a novel approach for predicting long-duration complex activity by discovering the causal relationships between constituent actions and predictable characteristic of the activities. The key of our approach is to utilize the observed action units as context to predict the next possible action unit, or predict the intension and effect of the whole activity. It is thus possible to make predictions with meaningful earliness and have the machine vision system provide a time-critical reaction. We represent complex activities as sequences of discrete action units, which have specific semantic meanings and clear time boundaries. To ensure a good discretization, we propose a novel temporal segmentation method for action units by discovering the regularity of motion velocities. We argue that the causality of action units can be encoded as Markov dependencies with various lengths, while the predictability can be characterized by a predictive accumulative function (PAF) learned from information entropy changes along every stage of activity progress.

Additionally, according to cognitive science, context information is critical for understanding human activities [34], [25], [28], [54], [21], [23], [10], which typically occur under particular scene settings with certain object interactions. So for activity prediction, it needs to involve not only actions, but also objects and their spatial temporal arrangement with actions. Such knowledge can provide valuable

- *K. Li is with the Department of Electrical and Computer Engineering, College of Engineering, Northeastern University, Boston, MA 02115. E-mail: li.ka@husky.neu.edu.*
- *Y. Fu is with the Department of Electrical and Computer Engineering, College of Engineering, and College of Computer and Information Science, Northeastern University, 403 Dana Research Center, 360 Huntington Avenue Boston, MA 02115. E-mail: yunfu@ece.neu.edu.*

Manuscript received 03 Mar. 2013; revised 22 Oct. 2013; accepted 06 Dec. 2013. Date of publication 1 Jan. 2014; date of current version 10 July 2014.
Recommended for acceptance by D. Forsyth.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2013.2297321

0162-8828 © 2013 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

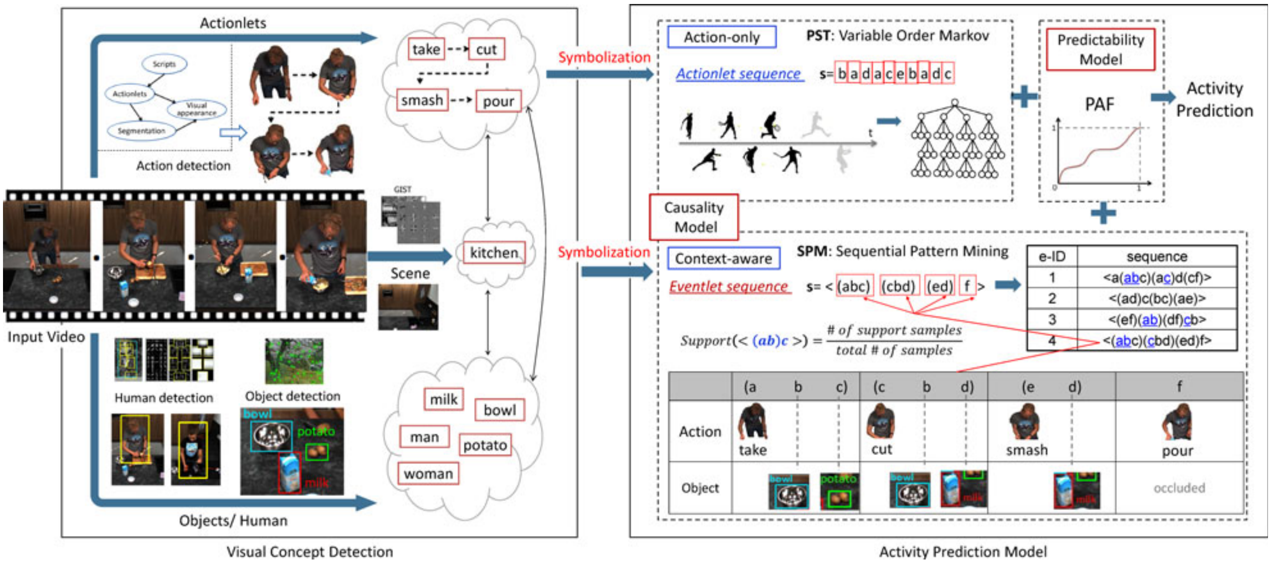


Fig. 1. Frameworks of long-duration complex activity prediction. Two scenarios: (1) Action-only activity prediction. (2) Context-aware activity prediction. The particular activity shown in the sample video is “making mashed potatoes”. The video data is from [43]. Our approach aims to solve activity prediction problem in both cases.

clues for two questions *‘what is happening now?’* and *‘what is goanna happen next?’*. Therefore, a unified approach is expected to provide unexplored opportunities to benefit from mutual contextual constraints among actions and objects. When a particular $\langle action, object \rangle$ pair is observed, the whole plan of human behavior may be inferred immediately. For example, as long as we observe *‘a person grabbing a cup’*, we probably can tell she/he is going to drink a beverage. In this paper, we extend our previous work [32], an action-only model, to include a context-aware model. We utilize sequential pattern mining (SPM) to incorporate the context information into actions which together can be represented as enriched symbolic sequences.

Overall, we propose a generalized activity prediction framework, which has four major components as shown in Fig. 1: (1) a visual concept detection module, (2) an action-only causality model, (3) a context-aware causality model, and (4) a predictability model. In order to test the efficacy of our method, evaluations were done on two experimental scenarios with two data sets for each: action-only prediction and context-aware prediction. Our method achieved superior performance for predicting global activity classes and local action units.

1.1 Related Work

In general, there are three categories of works that are mostly related to ours: complex activity recognition, early detection of actions or events,[1] and event prediction in AI.

Complex activity recognition. Recently, there has been a surge in interest in complex activity recognition by involving various structural information represented by spatial or temporal logical arrangements of several activity patterns. Most works aim to provide a good interpretation of complex activity. However, in many cases, inferring the goal of agents and predicting their plausible intended action are more desirable. Grammar based methods [24], [46] show effectiveness for composite human activity recognition. Pei et al. [41] proposed to deal with goal inference and intent prediction by parsing video events based on a Stochastic Context Sensitive Grammar (SGSG) which is automatically learned according to [47]. The construction of the hierarchical compositions of spatial and temporal relationships between the sub-events is the key contribution of their work. Without a formal differentiation between activity recognition and activity prediction, their system is actually doing an online detection of interesting events. Two important aspects for prediction, the earliness and the causality are missing in their discussion. The syntactic model is a very powerful tool for representing activities with high level temporal logic complexity. Hamid et al. [20] proposed the idea that global structural information of human activities can be encoded using a subset of their local event sequences. They regarded discovering structure patterns of activity as a feature selection process. Although rich temporal structure information was encoded, they did not consider prediction possibility from that point.

Although not directly dealing with activity prediction, several notable works present various ways to handle activity structure. Logic based methods are powerful in incorporating human prior knowledge and have a simple inference mechanism [6]. To model temporal structure of decomposable activities, Gaidon et al. [17] and Niebles et al. [39] extended the classic bag-of-words model by including segmentation and dynamic matching. Kwak et al. [29] and Fan et al. [16] regarded complex activity recognition as a constrained optimization problem. Wang et al. [52] introduced the actionlet ensemble model, in which spatial structures of the features were encoded.

Early detection of action/events. It is important to distinguish between early detection and prediction. Essentially they are dealing with prediction in different semantic

[1] Concepts “action” and “event” are always interchangeably used in computer vision and other AI fields. In our discussion, we prefer to use “action” when referring human activity, and use “event” to refer more general things, such as “stock rising.”

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

granularity. Early detection tries to recognize an ongoing atomic action from observation of its early stage. For example, an action of "handshaking" can be early detected by just observing "outstretched hand". However, for activity prediction, it tries to infer the intention or a higher level activity class with observation of only a few action units.

Ryoo [45] argued that the goal of activity prediction is to recognize unfinished single actions from observation of its early stage. Two extensions of bag-of-words paradigm, dynamic BoW and integral BoW are proposed to handle the sequential nature of human activities. Cao [9] extended Ryoo's work to recognize human activities from partially observed videos, where an unobserved subsequence may occur at any time by yielding a temporal gap in the video. Hoai and De la Torre [22] proposed a max-margin framework for early event detection, in which video frames are simulated as sequential event streams. Then a structure SVM based event detector is learned to recognize partially observed sequences. To deal with the sequential arrival of data, they developed a new monotonicity of detection function and formulated the problem within the structural support vector machine framework. Davis and Tyagi [12] presented a probabilistic framework for rapid detection of human actions. Their method achieves early detection by determining the shortest video exposures needed for a reliable recognition.

Event prediction in other fields. While human activity prediction has received little attention in the computer vision field, predicting events or agent behaviors have been extensively studied in many other AI fields. Neill et al. [38] proposed a new Bayesian method for prospective disease surveillance. Their method applied similar technique of abrupt change detection [13], and can predict disease locations resulting from emerging disease outbreaks. Brown et al. [8] described *n*-gram models for predicting the next words by using previous words as context in text samples. Event prediction also has been studied in the domain of Email-Spamming, where anti-spam techniques are applied to make immediate detection and prevention whenever a message arrives. By exploiting collective information about entire batches of spam messages, Haider et al. [19] proposed a method for detecting jointly generated spam emails effectively. Forecasting of financial time-series, such as stock index [26] predicts the future data based on the historical data. This type of techniques however are not suitable for activity prediction since it can only predict the micro value of next timestamp instead of recognizing the macro event classes. Kitani et al. [27] formulated the prediction task as a decision-making process [57] and proposed to forecast human behavior by leveraging the recent advances in semantic scene labeling [35], [36] and inverse optimal control [4], [1], [56], [31]. They predicted the destination of pedestrians and the routes they will choose.

1.2 Contributions of this work

In this section, aspects of our proposed activity prediction framework are highlighted:

- Our approach is the first to solve long-duration complex activity prediction problem, and has following theoretical contributions on modeling three key aspects of activity: 1) **Causality** by probabilistic suffix tree (PST), which can represent both large and small order Markov dependencies between action units; 2) **Context-cue** by sequential pattern mining, which utilizes interactive objects as cues for predicting human activity; and 3) **Predictability** by predictive accumulative function, which automatically learns the predictability pattern of each kind of activity from data.
- Our approach is a general framework for human activity prediction. 1) It can be integrated with any sequential decomposition methods of complex activity with flexible actionlets[2] granularity. 2) It can also include context information, such as objects, as strong cues for predicting high-level activity. 3) Useful observation derived from other sensors, besides cameras, can be easily added into the framework to improve prediction capability.
- The intuitions and corresponding models demonstrated by our approach should inspire new avenues of research on activity prediction.

2 Activity Prediction

To allow for more clarity in understanding our activity prediction model, we want to first provide an abstraction of activity prediction problem. Essentially, we transform the activity prediction problem into the problem of **early prediction on sequential data representation**. So the solution to this problem involves answering following two questions: 1) *"how to represent activity as a sequential data, which the way it is"*; 2) *"how to do early prediction on such kind of representation"*. We call the first one representation phase, and the second one prediction phase.

In the representation phase, an observation of complex human activity (e.g., from a camera, or from a rich networked sensor environments) is temporally segmented into semantic units in terms of component atomic actions (we call them actionlets). The boundaries between actionlets are detected by monitoring motion patterns (Section 3.1.1). Inside each segment, observed actionlets and objects are detected and quantized to symbolic labels which map to action and object classes.

In the prediction phase, the prediction problem becomes a sequence classification problem, but given only partial observation of the sequence (the beginning part). In data mining literature, sequence classification problem has three main categories of approaches: 1) **Feature selection** with traditional vector based classification. (e.g., *K*-grams as features); 2) **Distance function**, which measures similarity of two sequences. (KNN or SVM kernel can be used as classifier in this scenario). 3) **Model based method**, which simulates a generative process to get a sequence. A model trained on sequences in one class can assign a likelihood to a new sequence. Specific models include *K*-order Markov model, variable order Markov model (VMM), and hidden Markov model (HMM).

[2] In this paper, we use *actionlet* to refer to component *action units* obtained from temporal decomposition. These two terms are used interchangeably. It should be noted that the same term 'actionlet' has also been used in the recent work [52], which refers to action components based on a spatial segmentation.

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

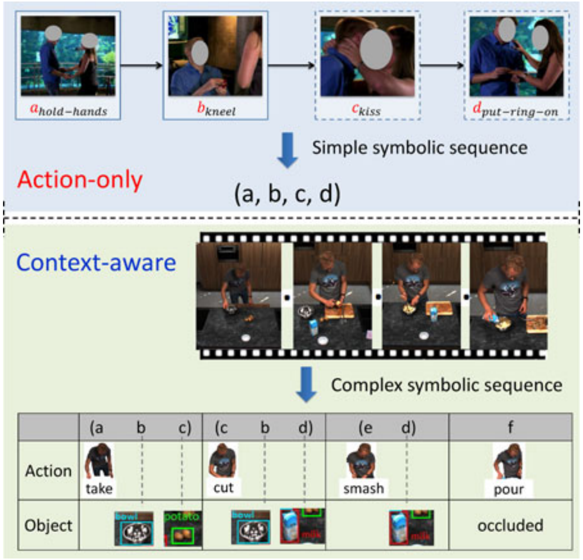


Fig. 2. Two scenarios of activity modeling: action-only model (top) and context-aware model (bottom).

Our approach proposes the prediction model in two scenarios: action-only and context-aware, which is characterized by what kind of information used for prediction. In both cases, we adopt the third strategy to train a model for prediction. The reasoning behind this choice is explained later. We formalize the representation first.

- Given an alphabet of actionlet symbols $\Sigma = \{a_1, a_2, a_3, \ldots, a_n\}$, an observation of activity is represented by a *simple symbolic sequence*, which is an ordered list of the actionlet symbols from the alphabet. An **action-only prediction model** takes this type of sequence of actionlets from an ongoing activity as input, and predicts the high level activity class. For example, the activity "marriage proposal" is composed of four actionlets: $\langle a_{hold-hands}, b_{kneel}, c_{kiss}, d_{put-ring-on}\rangle$, shown in Fig. 2 (top).
- Given an alphabet of semantic symbols (actionlet and object labels) $\Sigma = \{e_1, e_2, e_3, \ldots, e_n\}$. In **context-aware prediction model**, an observation of activity is represented by a *complex symbolic sequence*, which is an ordered list of vectors. Each vector is a subset of the alphabet. For example, the activity "cook smashed potato dish" is composed of following item sets: $\langle (abc)(cbd)(ed)f\rangle$, the meaning of each symbol is shown in Fig. 2 (bottom).

Now we specifically talk about our prediction models, and the reasons we chose them. For our action-only prediction model, we propose to use PST (an implementation of VMM) as our main causality model. The reason for using this model is that the first two categories of approaches: feature-selection-based and distance-function-based, can not handle partial sequence as input (i.e., useful patterns in the sequence that are highly dependent on the global observation). And among approaches in the third category, HMM can only model *1*-order dependency, so it will ignore a lot of long-term causality information between activity components, which we believe is essential for prediction. *K*-order Markov model restricts order number to a specific order, so it will lack flexibility, making it unable to include small order and large order causalities at the same time, or in an adaptive fashion. So VMM is the most suitable model for early classification of sequence data, which models long-duration and short-duration sequential dependency as causality, and requires no need to see the whole sequence.

For our context-aware prediction model, we propose to use sequential pattern mining to include objects cues for prediction. SPM is well suited for this problem because it uses item sets as sequence unit, so we can put co-occurrence of action and object as an enriched observation unit in complex activity scenario. Also, it can be easily tuned to fit into our whole prediction framework, which can be seen as an upgraded, rather compatible version of our action-only model. Details will be discussed in the later sections. Table 1 summarizes the capability of different methods.

2.1 Two Prediction Scenarios

To demonstrate the effectiveness of our proposed approach, we evaluate prediction tasks on two data sets for two different scenarios. For the action-only prediction, we test the ability of our approach to predict human daily activities, such as "making a phone call", which have middle-level temporal complexity. Next, we test our model at high-level temporal complexity activities on a tennis game data set collected by the authors. For the context-aware prediction, first we test our approach on a cooking activity data set, where the whole kitchen environment is observed from a static camera. Second, we extend our data domain to a networked sensor environments, where we want to predict several kinds of morning activities in an office lounge, such as "coffee time".

2.1.1 Action-Only Prediction Scenario

Our prediction model is applicable to a variety of human activities. The key requirement is that the activity should have multiple steps where each step constitutes a meaningful action unit. Without loss of generality, we choose two data sets with significant different temporal structure complexity. First, we collect real world videos for tennis games between two top male players from YouTube. Each point with an exchange of several strokes is considered as an activity instance, which involves two agents. In total, we

TABLE 1
Comparison of Different Methods in Terms of Capability for Modeling Activity Prediction

Models	Sequence classification	Partial sequence classification (prediction)	Causality	Context
Feature based	+	-	-	-
Distance function based	+	-	-	-
HMM	+	+	+ (of order 1)	-
Variable order Markov model	+	+	+	-
Sequential pattern mining	+	+	+	+

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

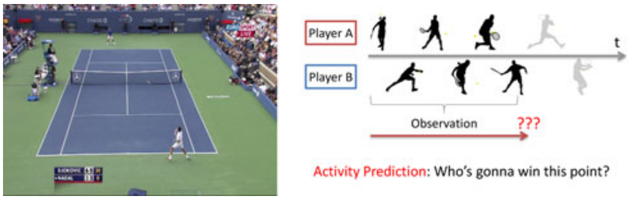


Fig. 3. Left: tennis game data set. Right: activity prediction task on this data set.

collected 160 video clips for 160 points from a 4 hour game. The clips were then separated into two categories of activity, where 80 clips are winning points and 80 clips are losing points with respect to each specific player. So our prediction problem on this data set becomes the question question: "can we predict who will win?". The data set and prediction task are illustrated in Fig. 3. Since each point consists of sequence of actionlets with length ranging from 1 to more than 20, tennis game has a high-level temporal structure complexity in terms of both variance and order.

Second, we choose the Maryland Human-Object Interactions (MHOI) data set [18], which consists of six annotated activities: *answering a phone call*, *making a phone call*, *drinking water*, *lighting a flash*, *pouring water into container* and *spraying*. These activities have about three to five action units each. Constituent action units share similar human movements: 1) reaching for an object of interest, 2) grasping the object, 3) manipulating the object, and 4) put back the object. For each activity, we have 8 to 10 video samples. There are 54 video clips in total. Examples in this data set are shown in Fig. 5.

2.1.2 Context-Aware Prediction Scenario

To verify our context-aware model, we perform experiments on two complex activity data sets, where human actions involve a lot of interactions with various objects.[3] The first is a fine-grained cooking activity data set, and the other is a complex morning activity data set in highly rich networked sensor environment.

The MPII cooking activities data set (MPII-Cooking) [43] contains 44 instances of cooking activity, which are continuously recorded in a realistic setting, as shown in Fig. 4 (top). Predictable high level activities are about preparing 14 kinds of dishes, including: *making a sandwich*, *making a pizza*, and *making an omelet*, etc. There are overall 65 different actionlets as building blocks shared among various cooking activities, such as *cut*, *pour*, *shake*, and *peel*.

The OPPORTUNITY activity recognition data set (UCI-OPPORTUNITY) [42] was created in a sensor-rich environment for the machine recognition of human activities, as shown in Fig. 4 (bottom). They deployed 72 sensors of 10 modalities in 15 wireless and wired networked sensor systems in the environment, on the objects, and on the human body. The data are acquired from 12 subjects performing morning activities, yielding over 25 hours of sensor data. It contains five high level predictable activities (*Relaxing*, *Coffee time*, *Early Morning*, *Cleanup*, *Sandwich time*), 13 low level actionlets (e.g., *lock*, *stir*, *open*, *release*), and 23 interactive objects (e.g., *bread*, *table*, *glass*).

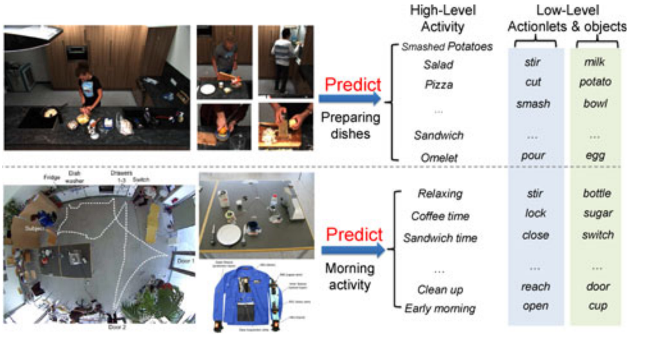


Fig. 4. Two data sets for context-aware prediction scenario. Top: MPII-Cooking data set [43], where we want to predict the type of dish the subject is preparing. Bottom: UCI-OPPORTUNITY data set [42], where we want to predict current ongoing morning activity.

3 Proposed Approach

3.1 Representation: Activity Encoding

3.1.1 Actionlets Detection by Motion Velocity

Temporal decomposition is the first key step for our representation of complex activity. It is used to find the frame indices that can segment a long sequence of human activity video into multiple meaningful atomic actions. Relevant work can be found in [50]. We call these atomic actions **actionlets**. We found that the velocity changes of human actions have similar periodic regularity. Fig. 5 shows three examples of actionlet segmentation and detection.

The specific method includes these steps: 1) use Harris corner detector to find significant key points; 2) use Lucas-Kanade (LK) optical flow to generate the trajectories for key points; 3) for each frame, accumulate the trajectories/tracks at these points to get a velocity magnitude:

$$V_t = \sum_{p(x_{i,t}, y_{i,t}) \in F_t} \sqrt{(x_{i,t} - x_{i,t-1})^2 + (y_{i,t} - y_{i,t-1})^2}, \qquad (1)$$

where V_t represents the overall motion velocity at frame F_t, p_i is the ith interest point found in frame F_t. $(x_{i,t}, y_{i,t})$ is the position of point p_i in the frame. We observed that each hill in the graph represents a meaningful atomic action. For each atomic action, the start frame and the end frame always have the lowest movement velocity. The velocity reaches the peak at the intermediate stage of each actionlet. To evaluate our temporal decomposition approach, a target window with the size of 15 frames around the human labeled segmentation point is used as the ground truth. We manually labeled 137 segmentation points for all 54 videos in the MHOI data set. The accuracy of automatic actionlets segmentation is 0.83. For the tennis game data set, we cut the video clips into top-half and bottom-half to handle actionlets of two players. We labeled 40 videos with 253 actionlets in it. The actionlet segmentation accuracy is 0.82.

3.1.2 Activity Encoding

Based on accurate temporal decomposition results, we can easily cluster actionlet into meaningful groups so that each

[3] Data sets selection is based on two criterions: 1) the object information should not dominate the recognition of activity; 2) the interactive objects should contribute extra semantic information to actionlets.

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

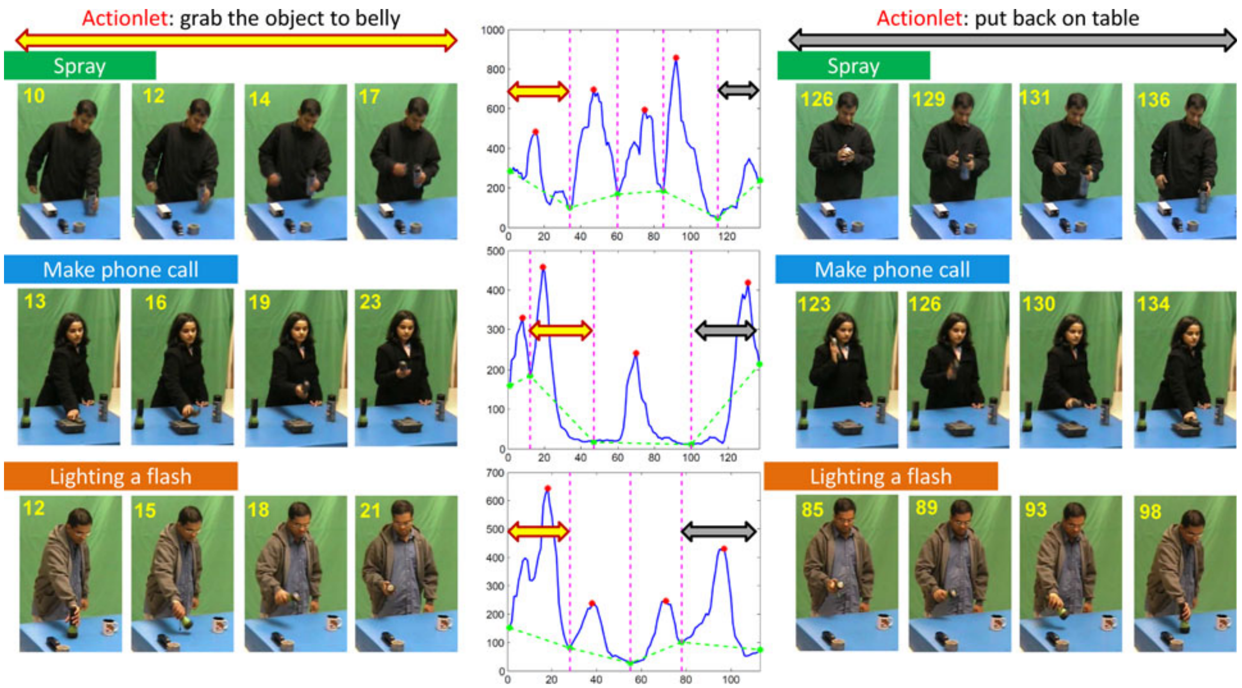


Fig. 5. Actionlet detection by motion velocity. Time series figures show motion velocity changes over time (smoothed). The horizontal axis is the frame index, and the vertical axis is the velocity strength computed according to formula (1). Red dots on each local peak indicate actionlets centers, and green dots on each local valley indicate the segmentation points. A point is considered an actionlet center only if it has the local maximal value, and was preceded (to the left) by a value lower by a threshold. Actionlets are obtained by extracting segments between two consecutive vertical dashed lines. Two actionlets shared by three different types of actions are shown as examples: "*grab the object to belly*" (left) and "*put back on the table*" (right). Images on each row are from the same video marked with frame indices.

activity can be represented by a sequence of actionlets in a syntactic way. A variety of video descriptors can be used here as long as they can provide discriminative representations for the actionlets.

Due to different spatial extent of humans in the scene and different background motion styles, two approaches are used to compute descriptors for tennis game data set and MHOI data set respectively. For the MHOI data set which has a large scale human in the scene and a static background, we use the 3-D Harris corner detector to find sparse interest points. Each local area is described by Histogram of Gradients (HoG) and Histogram of Flow (HoF) descriptors [30]. Furthermore, we vector quantize the descriptors by computing memberships with respect to a descriptor codebook of size 500, which is obtained by k-means clustering of the descriptors. Then, actionlets categories are learned from histogram of spatial-temporal words using an unsupervised algorithm [40]. To evaluate the actionlets encoding results, human expert watch video segments corresponding to each actionlet, and annotate them according to their semantic meanings, such as "*reach the object*", "*grab the object to belly*" and "*grab the object to head*" etc. The Rand index[4] of clustering is 0.81.

For the tennis game data set, the scale of player in the video is very small, therefore it is difficult to get sufficient local features by using sparse sampling methods. Here, we use dense trajectories [51] to encode actionlets. For every actionlet, we sample the initial feature points every w pixels at multiple spatial scales. All tracking points are obtained by a median filter in a dense optical flow field from the points in the previous frame. For each trajectory, the descriptor is calculated in a 3-D volume. Each such volume is divided into sub-volumes. HOG, HOF and MBH features are then computed for every sub-volume. In our approach, we use the same parameters indicated in [51]. The codebook size we used is 1,000. In addition, to remove the noises caused by camera movements and shadows, a human tracker [11] is used before extracting feature records. For evaluation, we group 253 actionlets from 40 annotated videos into 10 categories, and the Rand index of clustering is 0.73. For MPII-Cooking and UCI-OPPORTUNITY data sets, we use pre-annotated actionlet boundaries provided by the data set when we perform activity encoding. Our focus on these two data sets is to evaluate context-aware prediction model.

3.2 Action-Only Causality Model

Here we introduce the model of human activity prediction, which is illustrated in Fig. 1. Let Σ be the finite set of actionlets, which are learned from videos using unsupervised segmentation and clustering methods. Let $D_{\text{training}} = \{r^1, r^2, \dots, r^m\}$ be the training sample set of m sequences over the actionlet alphabet Σ, where the length of the ith $(i = 1, \dots, m)$ sequence is l_i (i.e., $r^i = r^i_1 r^i_2 \dots r^i_{l_i}$, where $r^i_j \in \Sigma$). Based on D_{training}, the goal is to learn a model P that provides a probability assignment $p(t)$ for an ongoing actionlet sequence

[4] Rand index is a measure of the similarity between data clustering and ground truth. It has a value between 0 and 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same.

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

$t = t_1, t_2, \ldots, t_{\|t\|}$. To realize this design with maximum predictive power, we include two sources of information in the model. One is the causality cue hidden in the actionlet sequences, which encodes the knowledge about the activity. The other is the unique predictable characteristic for each kind of human activity, which answers the questions why a particular activity can be predicted and how early an activity can be predicted with satisfactory accuracy.

Algorithm 1 Construction of L-bounded PST $\overline{T}$ $(L, P_{\min}, \alpha, \beta, \lambda)$

1) **Forming candidate suffix set** $\overline{S}$: Let $D_{\text{training}} = \{r^1, r^2, \ldots, r^m\}$ be the training set, and assume s is a subsequence of r^i $(i = 1, \ldots, m)$. If $|s| < L$ and $P(s) > P_{\min}$, then put s in $\overline{S}$. $P_{\min}$ is a user specified minimal probability requirement for an eligible candidate. $P(s)$ is computed from frequency count.

2) **Testing every candidate** $\mathbf{s} \in \overline{\mathbf{S}}$: For any $s \in \overline{S}$, test following two conditions:
   - (1) $P(\sigma|s) \geq \alpha$, which means the context subsequence s is meaningful for some actionlet σ. Here, α is defined by user to threshold a conditional appearance.
   - (2) $\frac{P(\sigma|s)}{P(\sigma|\text{suf}(s))} \geq \beta$, or $\leq 1/\beta$, which means the context s provides extra information in predicting σ relative to its longest suffix $\text{suf}(s)$. β is a user specified threshold to measure the difference between the candidate and its direct parent node.
   - **Then**, if s passes above two tests, add s and its suffixes into $\overline{T}$.

3) **Smoothing the probability distributions to obtain** $\gamma_{\mathbf{s}}(\sigma)$:
   For each s labeling a node in $\overline{T}$, if $P(\sigma|s) = 0$, we assign a minimum probability λ. In general, the *next symbol probability function* can be written as:
   $\gamma_s(\sigma) = (1 - |\Sigma|\lambda)P(\sigma|s) + \lambda$. Here, λ is the smoothing factor defined empirically.

Causality is an important cue for human activity prediction. Our goal is to automatically acquire the causality relationships from sequential actionlets. Variable order Markov model [5] is a category of algorithms for prediction of discrete sequences. It suits the activity prediction problem well, because it can capture both large and small order Markov dependencies extracted from training data. Therefore, it can encode richer and more flexible causal relationships. Here, we model complex human activity as a probabilistic suffix tree [44] which implements the single best L-bounded VMM (VMMs of degree L or less) in a fast and efficient way.

The goal of the PST learning algorithm is to generate a conditional probability distribution $\gamma_s(\sigma)$ to associate a "meaningful" context $s \in \Sigma^{\star}$ with the next possible actionlet $\sigma \in \Sigma$. We call the function $\gamma_s(\sigma)$ the *next symbol probability function*, and denote the trained PST model as $\overline{T}$, with corresponding suffix set as $\overline{S}$ consisting of actionlets sequence of all the nodes. Algorithm 1 shows the detailed building process of PST, where there are five user specified parameters. Fig. 6 shows an example of PST constructed from a training sequence of actionlets.

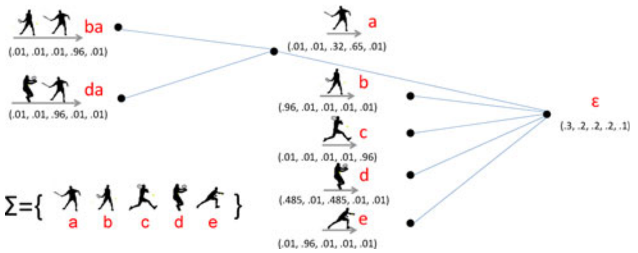


Fig. 6. Example of PST corresponding to the training sequence $r = badacebadc$ over alphabet $\Sigma = \{a, b, c, d, e\}$. The vector under each node is the probability distribution over alphabet associated with the actionlets subsequence (in red). (e.g., the probability to observe d after a subsequence, whose largest suffix in the tree is ba, is 0.96).

3.3 Context-Aware Causality Model

The context-aware causality model embodies vocabularies of visual elements including actionlets and objects as enriched symbolic sequences that specify syntactic (compositional) relationships. We call each unit of the enriched symbolic sequence eventlet.[5] Sequential pattern mining [33], [55] was first introduced in the area of data mining, where a sequence database stores a set of records consisting of sequences of ordered events, with or without concrete notions of time. SPM algorithm mines the sequence database looking for repeating patterns (known as frequent sequences) that can be used later by the end-users to find associations between different items or events in their data for purposes such as marketing campaigns, web usage mining[15], [37], [49], DNA sequences analysis [53], prediction and planning. There are two main reasons we propose to use SPM as our context-aware model. First, our framework detects a sequence of visual concepts, which are well quantized as semantical labels with concrete time stamps. Essentially the data structure we are dealing with is quite similar to the common data types in the sequential database, such as customer transactions data or purchase sequences in a grocery store. Secondly, in our context-aware scenario, actionlets have been enriched with co-occurred objects, so the newly formed eventlet sequence has both the compositional and sequential nature, which can be perfectly fitted into a SPM model.

We present two algorithms, which we call *Mapping-based (Algorithm 2)* and *SPM-based (Algorithm 3)*. The *mapping-based* algorithm is a simplified version of our context-aware causality model, which transforms the complex symbolic sequence (eventlets) into a simple symbolic sequence. Then a similar PST model can be applied for the mapped simple symbolic sequences, as we used in the action-only model.

[5] In this paper, we use eventlet to refer to observation of actionlet and objects co-occurrence. An eventlet $e = \langle \{a^*\} \bigcup \{o_1, o_2, \ldots, o_m\} \rangle$, where a^* represents a particular actionlet, and o_i represents a particular object interacting with a^* within its segment. In our case, n will always be 0, 1, or 2 with the meaning of none, one, or two co-occurrent interacting objects (we assume one person at most can operate two different objects at the same time with two hands).

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
Transformed Representation: Mapping Complex Symbolic Sequence into Simple Symbolic Sequence

Activity Instance ID	Original Sequence	Frequent Itemset and mapping	After Mapping	After Transformation
1	$\langle a, b\rangle$	$a \to 1, b \to 5$	$\langle 1, 5\rangle$	$\langle 1, 5\rangle$
2	$\langle (cd), a, (efg)\rangle$	$a \to 1, e \to 2, g \to 3, (eg) \to 4$	$\langle 1, \{2, 3, 4\}\rangle$	$\langle 1, 2\rangle$, $\langle 1, 3\rangle$, $\langle 1, 4\rangle$
3	$\langle (ahg)\rangle$	$a \to 1, g \to 3$	$\langle \{1, 3\}\rangle$	$\langle 1\rangle$, $\langle 3\rangle$
4	$\langle a, (eg), b\rangle$	$a \to 1, e \to 2, g \to 3, (eg) \to 4, b \to 5$	$\langle 1, \{2, 3, 4\}, 5\rangle$	$\langle 1, 2, 5\rangle$, $\langle 1, 3, 5\rangle$, $\langle 1, 4, 5\rangle$
5	$\langle b\rangle$	$b \to 5$	$\langle 5\rangle$	$\langle 5\rangle$

Algorithm 2 Mapping-based Context-aware model

1) **Frequent Itemsets Phase**:
   Given ξ_{itemset}, find the set of all frequent itemsets FI applying apriori algorithms [3].
   Notice: A small difference here is that the support of an itemset is defined as the fraction of activity instances (the sequences of eventlets) rather the fraction of eventlets.
2) **Mapping Phase**:
   $f : FI \to FI'$, where $FI' \subset Z$, frequent itemset $i_s \in FI$ is mapped into an integer $x \in FI'$.
3) **Transformation Phase**: FI' is further broken down into individual interactions of frequent itemsets along the time line, e.g. the examples in Table 2.
   Notice: The reason for this mapping is that by treating frequent itemsets as single entities, we can transform our context-aware observation, a complex symbolic sequence, into a simple symbolic sequence representation. This transformed representation is called D_{training}.
4) **Construct Causality Model**: Use the set D_{training} to build the causality model by calling Algorithm 1.

The *SPM-based* algorithm is a relatively complex version of our context-aware causality model, which finds frequent subsequence of item sets as sequential patterns. Then we utilize the mined sequential patterns to compute the conditional probabilities $\gamma_s(\sigma)$, which associates a "meaningful" context $s \in \Sigma^\star$ with next possible eventlet. In our context-aware model as shown in Fig. 1, we have following definitions.

- A set of observations of human activity, saying activity archive D, is represented as a set of sequential records of eventlets. Each eventlet is segmented according to actionlet boundaries, and represented as an item set of detected visual concepts.
- An eventlet sequence is an ordered list of item sets, for example, $s = \langle a(be)c(ad)\rangle$. An item set is a set drawn from items in Σ, and denoted $(i_1, i_2, \ldots, i_k)$, such as a and (be) in the previous example. Σ is a set of N unique items $\Sigma = i_1, i_2, \ldots, i_N$, where each element in Σ can be either an actionlet label or an object label.
- The support of item set or eventlet, $i_s = (i_1, i_2, \ldots, i_k)$ is defined as the fraction of activity instances (the sequences of eventlets) $s \in D$ that contains the item set in any one of its possibly many eventlets. Given a support threshold $\text{min\_sup}\ \xi_{\text{itemset}}$, an item set is called frequent item set on D if $\text{sup}_D(i_s) \geq \xi_{\text{itemset}}$. The set of all frequent item sets is denoted as FI.
- The support of a sequence s_a is defined as the fraction of activity instances $s \in D$ that contains s_a, denoted by $\text{sup}_D(s_a)$. Given a support threshold $\text{min\_sup}\ \xi_{sequence}$, a sequence is called a frequent sequential pattern on D if $\text{sup}_D(s_a) \geq \xi_{sequence}$.
- The length of a sequence is the number of item sets in the sequence. SP_k denotes the set of frequent sequential pattern with length k. So $SP_1 = FI$.
- Either based on Algorithm 2 or Algorithm 3, we can now train a contex-aware causality model $\gamma_s(\sigma)$, as we did in the action-only case.

Algorithm 3 SPM-based Context-aware model

1) **Frequent Itemsets Phase**:
   Given ξ_{itemset}, find the set of all frequent itemsets FI applying Apriori algorithms [3].
2) **Mapping Phase**:
   $f : FI \to FI'$, where $FI' \subset Z$, frequent itemset $i_s \in FI$ is mapped into an integer $x \in FI'$.
3) **Sequential Pattern Mining Phase**:
   Initialize $SP_1 = FI'$. Use algorithm AprioriAll[2] to find $SP_2, \ldots, SP_k$, where k is the largest length of frequent sequential pattern.
4) **Construct Causality Model**: Based on mined $SP_1, \ldots, SP_k$ and corresponding support for each sequential pattern in it, causality model $P(\sigma|s)$ can be computed by Bayes rules $P(\sigma|s)$, assume $|s| = k'$.
   - If $s\sigma \in SP_{k'+1}$, $P(s\sigma) = \frac{\text{sup}_D(s\sigma)}{\sum_{x_i \in SP_{k'+1}} \text{sup}_D(x_i)}$, $P(s) = \frac{\text{sup}_D(s)}{\sum_{x_i \in SP_{k'}} \text{sup}_D(x_i)}$, $P(\sigma|s) = \frac{P(s\sigma)}{P(s)}$.
   - Otherwise, $P(\sigma|s) = 0$.
5) **Smoothing the probability distributions to obtain** $\gamma_{\mathbf{s}}(\sigma)$:
   For each s, if $P(\sigma|s) = 0$, we assign a minimum probability λ. In general, the *next symbol probability function* can be written as:
   $\gamma_s(\sigma) = (1 - |\Sigma|\lambda)P(\sigma|s) + \lambda$. Here, λ is the smoothing factor defined empirically.

3.4 Predictive Accumulative Function

In this section, we want to answer "why a particular activity can be predicted", and how to make our model automatically adapted to activities with different predictability. For example, "tennis game" is a late-predictable problem in the sense that a long sequence of actionlets performed by two players was observed, the last several strokes will strongly impact the winning or losing results. In contrast, "drinking water" is an early predictable problem, since as long as we observed the first actionlet "grabbing a cup", we probably can guess the intention. To characterize the predictability of activities, we formulate a predictive accumulative function. Different activities usually reflect very different PAFs. In our model, PAF can be learned automatically from the training data.

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

Later when do prediction, we use PAF to weight the observed patterns in every stage of ongoing sequence.

Suppose $k \in [0,1]$ indicates the fraction of beginning portion (prefix) of any sequence. D is the training set. Let D_k be the set of sequences, where each sequence consists of the first k percentage of the corresponding $r = (r_1, r_2, \ldots, r_l) \in D$, where $r_i (i = 1, 2, \ldots, l) \in \Sigma$, l is the length of r. We use $r_{pre(k)}$ to represent the corresponding "prefix" sequence of r in D_k. Obviously $|D| = |D_k|$.

Given the first k percentage of the sequence observed, the information we gain can be defined as following:

$$y_k = \frac{H(D) - H(D|D_k)}{H(D)}. \tag{2}$$

Here the entropy $H(D)$ evaluates the uncertainty of a whole sequence, when no element is observed, and the conditional entropy $H(D|D_k)$ evaluates the remaining uncertainty of a sequence after first k percentage of sequence are checked:

$$\begin{aligned} H(D) &= -\sum_{r\in D} p^{\overline{T}}(r) \log p^{\overline{T}}(r), \\ H(D|D_k) &= -\sum_{r_{\mathrm{pre(k)}} \in D_k} \sum_{r \in D} p^{\overline{T}}(r, r_{\mathrm{pre(k)}}) \log p^{\overline{T}}(r|r_{\mathrm{pre(k)}}). \end{aligned} \tag{3}$$

Since $r_{\mathrm{pre(k)}}$ is the "prefix" of r, we have

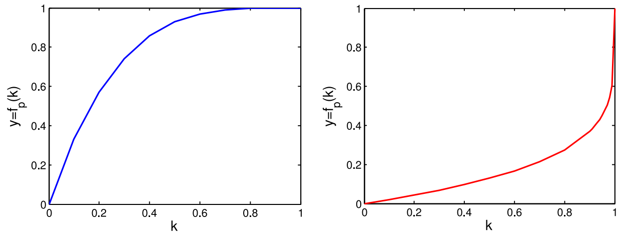
$$\begin{aligned} & p^{\overline{T}}(r, r_{\mathrm{pre(k)}}) = p^{\overline{T}}(r), \text{ and} \\ & p^{\overline{T}}(r|r_{\mathrm{pre(k)}}) = \frac{p^{\overline{T}}(r, r_{\mathrm{pre(k)}})}{p^{\overline{T}}(r_{\mathrm{pre(k)}})} = \frac{p^{\overline{T}}(r)}{p^{\overline{T}}(r_{\mathrm{pre(k)}})}. \end{aligned} \tag{4}$$

From trained PST model $\overline{T}$, we write

$$\begin{aligned} p^{\overline{T}}(r) &= \prod_{j=1}^{\|r\|} \gamma_{s^{j-1}}(r_j), \text{ and} \\ p^{\overline{T}}(r_{\mathrm{pre(k)}}) &= \prod_{j=1}^{\|r_{\mathrm{pre(k)}}\|} \gamma_{s^{j-1}}(r_{\mathrm{pre(k)}_j}). \end{aligned} \tag{5}$$

The nodes of $\overline{T}$ are labeled by pairs (s, γ_s), where s is the string associated with the walk starting from that node and ending in tree root; and $\gamma_s : \Sigma \rightarrow [0,1]$ is the *next symbol probability function* related with s, $\sum_{\sigma\in\Sigma} \gamma_s(\sigma) = 1$.

Based on above discussions, we can have a sequence of data pair (k, y_k) by sampling $k \in [0,1]$ evenly from 0 to 1. For example, by using 5 percent as interval, we will collect 20 data pairs. Now we can fit a function f_p between variable k and y, which we call predictive accumulative function: $\mathbf{y} = \mathbf{f_p}(\mathbf{k})$. Function f_p depicts the predictable characteristic of a particular activity. Fig. 7 shows PAFs in two extreme cases. The curves are generated on simulated data to represent an early predictable problem and a late predictable problem respectively.



(a) Early predictable problem.

(b) Late predictable problem.

Fig. 7. PAFs for depicting predictable characteristics of different activities.

3.5 Final Prediction Model

Given an ongoing sequence $t = t_1, t_2, \ldots, t_{\|t\|}$, we can now construct our prediction function by using the knowledge learned from Section 3.2/3.3 (causality) and Section 3.4 (predictability):

$$p^{\overline{T}}(t) = \sum_{j=1}^{\|t\|} f_p\left(\frac{\|t_1 t_2 \ldots t_j\|}{\|t\|}\right) \log \gamma_{s^{j-1}}(t_j), \tag{6}$$

which computes the weighted log-likelihood of t as the prediction score with the knowledge of trained PST model $\overline{T}$ and learned PAF f_p.

Giving an observed ongoing sequence of actionlets, our ultimate goal is to predict the activity class it belongs to. This problem can fit into the context of supervised classification where each class $c (c = 1, \ldots, C)$ is associated to a prediction model $p^{\overline{T}_c}(t)$ for which the empirical probabilities are computed over the whole set of sequences of this class belonging to the training set. Given an ongoing sequence $t = t_1, t_2, \ldots, t_{\|t\|}$, the sequence t is assigned to the class $c0$ corresponding to the prediction model $p^{\overline{T}_{c0}}$ for which maximal prediction score has been obtained: $\mathbf{p}^{\overline{\mathbf{T}}_{\mathbf{c0}}}(\mathbf{t}) = \mathrm{Max}\{\mathbf{p}^{\overline{\mathbf{T}}_{\mathbf{c}}}(\mathbf{t}), \mathbf{c} = \mathbf{1}, \ldots, \mathbf{C}\}$.

4 Experiments

We present experimental results on two scenarios of activity prediction: action-only and context-aware.

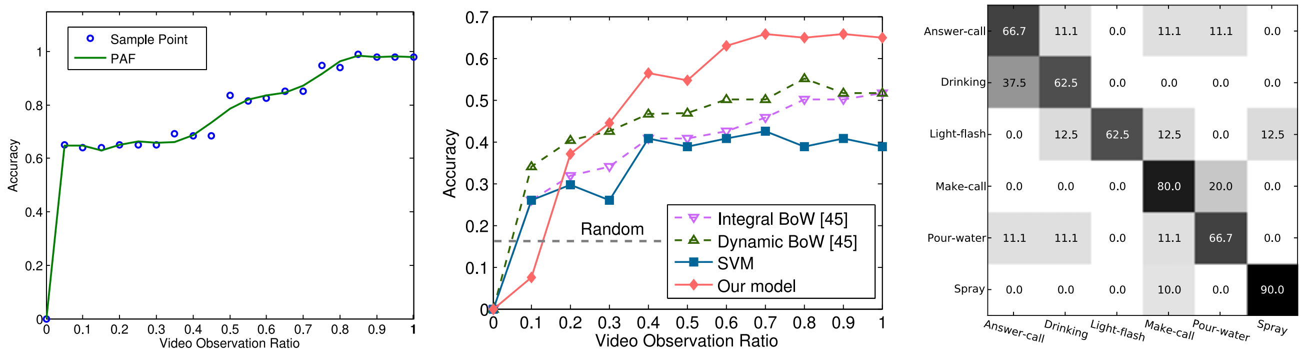
4.1 Action-Only Activity Prediction

4.1.1 Middle-Level Complex Activity Prediction

Samples in MHOI data set are about daily activities (e.g., "making phone call"). This type of activity usually consists of three to five actionlets and lasts about 5 to 8 seconds, so we call it middle-level complex activity. In this data set, each category has 8 to 10 samples. We evaluate the prediction accuracy by using the standard "leave-one-out" method, and fit activity prediction in the context of multi-class classification problem. Different from traditional classification task, for activity prediction, we focus on the predictive power of each method. The goal is to use an observation ratio as small as possible to make an accurate prediction. To train a prediction model, we constructed an order five-bounded PST and fit a PAF respectively. We compare our method of activity prediction with existing alternatives, including: (1) Dynamic bag-of-words model [45], (2) integral bag-of-words model [45], and (3) a basic SVM-based approach. All baseline methods are adopting the same bag-of-words representation with a codebook size of 500, which is built from the local features HOG/HOF.

Fig. 8a illustrates the process of fitting PAF from training data. It shows that daily activities such as

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.



(a) Fitting PAF. (b) Daily activity prediction. (c) Confusion matrix at 60% observation ratio.

Fig. 8. Activity prediction results on MHOI data set. (a) shows PAF of daily activity. (b) shows comparison of prediction accuracy of different methods. A higher curve indicates better performance of the corresponding method in terms of both prediction accuracy and earliness. Our approach shows the best performance. (c) shows the confusion matrix at 60 percent of observation.

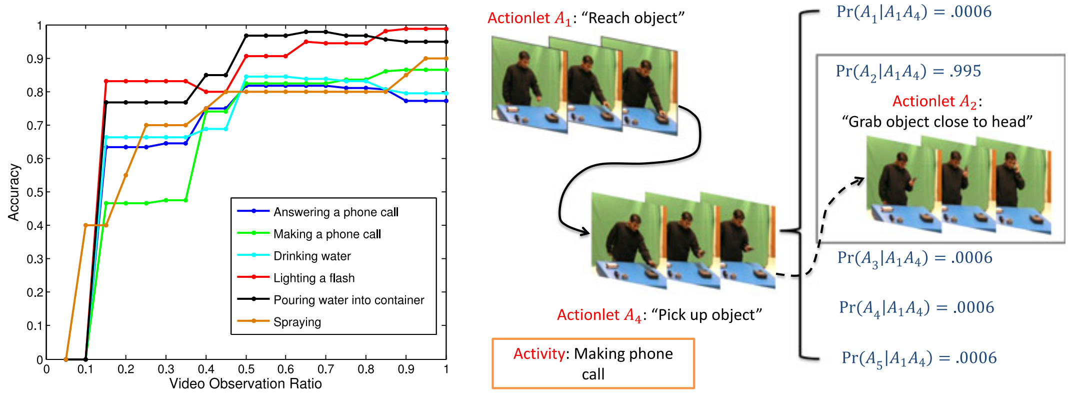
examples from MHOI data set are early predictable. That means the semantic information at early stage strongly exposes the intension of the whole activity. Fig. 8b illustrates the performance curves of the implemented four methods. The results are averaged over six activities. Its horizontal axis corresponds to the observed ratio of the testing videos, while the vertical axis corresponds to the activity recognition accuracy. The figure confirms that the proposed method has great advantages over other methods. For example, after half of the video is observed (about two actionlets), our model is able to make a prediction with the accuracy of 0.6.

Fig. 9a shows detailed performance of our approach over six different daily activities in a binary classification setting. From the figure, we can see that the activity "Pouring water into container" has the best prediction accuracy and earliness. In this data set, after the actors reach the object, they usually grab the object and put it close to the head. Three activities ("making a phone call", "drinking water", and "answering a phone call") share this process in the initial phase of the activity. So, in the activity "pouring water into container", after the first common actionlet "reach object", the second and third constituent actionlets make the sequence pattern quite distinctive. Besides predicting global activity classes, our model can also make local predictions. That means given observed actionlet sequence as context, the model can predict the most probable next actionlet. Fig. 9b shows an example from our experiment results.

4.1.2 High-Level Complex Activity Prediction

In this experiment, we aim to test the ability of our model to leverage the temporal structure of human activity. Each sample video in the tennis game data set is corresponding to a point which consists of a sequence of actionlets (strokes). The length of actionlet sequence of each point can vary from 1 to more than 20. So the duration of some sample videos may be as long as 30 seconds. We group samples into two categories, winning and losing, with respect to a specific player. Overall, we have 80 positive and 80 negative samples respectively. Then a six-bounded PST and a PAF are trained from data to construct the prediction model. The same "leave-one-out" method is used for evaluation.

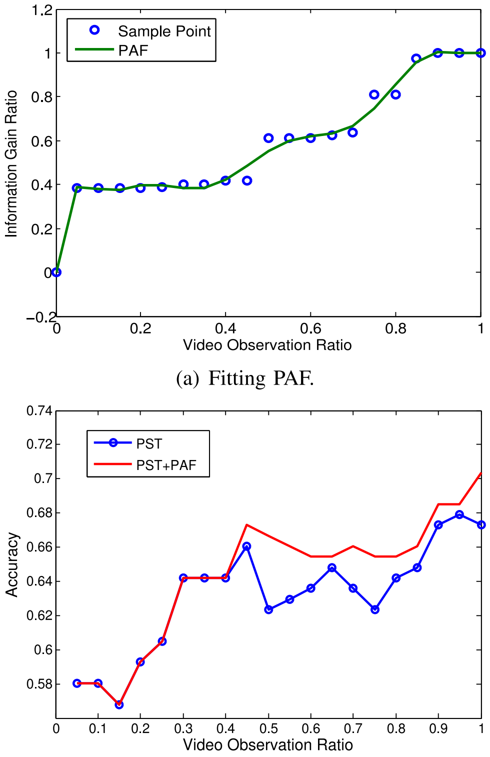
Fig. 10a illustrates the fitted PAF for tennis activity. It shows that tennis games are late predictable. That means the semantic information at late stage strongly impacts the results of classification. This is consistent with common sense about tennis games. Fig. 10b shows prediction performance of our method. Here we compare two versions of our model to illustrate the improvement caused



(a) Prediction on 6 activities. (b) Predict next actionlet.

Fig. 9. Global and local prediction for a particular activity in MHOI data set.

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.



(a) Fitting PAF.

(b) Tennis game prediction.

Fig. 10. Activity prediction results on tennis game data set. (a) shows PAF of tennis game. (b) shows prediction performance of our model. We did not show comparisons with the other three methods because of their inability to handle high-level complex activity, such as tennis game. Their prediction curves are nearly random, since bag-of-words representation is not discriminative anymore in this situation.

by considering predictable characteristic of activity. Since all other three methods, D-BoW, I-BoW and SVM, failed in prediction on this data set, we did not show comparisons. In short, our model is the only one that has the capability to predict on high-level complex activity. Table 3 shows detailed comparisons of seven methods on two data sets, where we include three sequence classification methods with each one representing a category of approaches. The details about other three methods will be discussed in Section 3.2.

4.1.3 Model Parameters

Advantage of our approach is that there are very few model parameters need to be tuned. Among them, the order L of PST is the most important one, since it determines the order of causal relationships that we want to incorporate. Fig. 11 shows the impact of parameter selections on prediction accuracy. We can see higher order of PST performs better. This is because it includes long-term and short-term Markov dependencies at the same time.

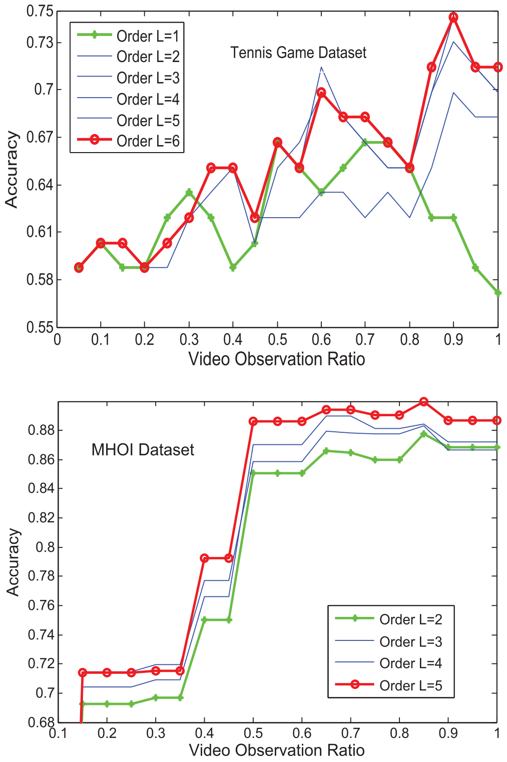


Fig. 11. Comparison of Different Order Bounded-PSTs. Left: prediction performance on Tennis Game data set. Right: prediction performance on MHOI data set.

4.2 Context-Aware Activity Prediction

In this section, we report experimental results on context-aware situations, which demonstrate that presence of interactive objects may significantly improve activity prediction performance by providing discriminative contextual cues often at very early stage of the activity progress.

4.2.1 Experiments on Cooking Activity Data Set

Samples in MPII-Cooking data set are about preparing various dishes (e.g., *making a sandwich*). Because of the varying degrees of complexity of different dishes, the length of eventlets sequence of each sample can vary from 20 to more

TABLE 3
Performance Comparisons on Two Data Sets

Methods	Tennis Game Dataset					MHOI dataset				
	20%	40%	60%	80%	100%	20%	40%	60%	80%	100%
Integral BoW [45]	0.52	0.49	0.47	0.48	0.44	0.32	0.41	0.42	0.50	0.52
Dynamic BoW [45]	0.53	0.51	0.47	0.49	0.53	**0.40**	0.47	0.50	0.55	0.52
BoW+SVM	0.56	0.52	0.51	0.48	0.49	0.30	0.41	0.41	0.39	0.39
Feature-based Model (K-grams) [14]	0.45	0.54	0.51	0.47	0.43	0.33	0.38	0.40	0.47	0.49
Distance Function base Model [48]	0.52	0.56	0.48	0.49	0.48	0.14	0.27	0.51	0.53	0.53
HMM	0.43	0.46	0.51	0.46	0.58	0.23	0.38	0.56	0.47	0.43
Our Action-only Model	**0.59**	**0.64**	**0.65**	**0.65**	**0.70**	0.37	**0.57**	**0.63**	**0.65**	**0.65**

Random guess for MHOI data set and Tennis Game data set are 0.167 and 0.5 respectively. Actually comparison methods perform random guess on tennis game. (Percentage as observation ratios).

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

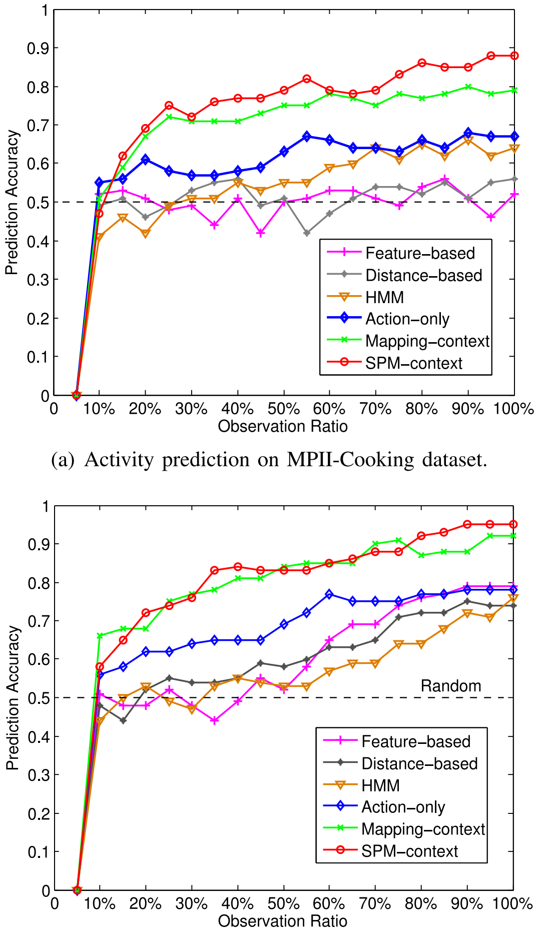
than 150. The average sequence length is 67.[6] For a particular activity, similar to experimental settings in action-only situation, we use all the samples in that category as the set of positive examples, and randomly select equal number of samples from remaining categories as the set of negative examples. Then we use the prediction task in the context of supervised binary classification of sequence data with varying observation ratios. To train a mapping-based context-aware model, we set $\xi_{\text{itemset}} = 0.3$, and put the transformed representation into a 11-order bounded PST to obtain a causality model. The PAF is also generated from the transformed representation according to the same method as before. To train a SPM-based context-aware model, we set $\xi_{\text{itemset}} = 0.3$ and $\xi_{\text{sequence}} = 0.2$.

To show the advantages of the proposed method, we compare our results with other three sequence classification methods, which may represent each of the three main categories of approaches. The details of comparison methods are as follows.

- ***k*-gram** ($k = 2$) with linear-SVM [14] represents feature-base methods. *k*-gram is the most popular feature selection approach in symbolic sequence classification.
- The **Smith-Waterman algorithm** with KNN [48] represents sequence distance-based methods. Since in our prediction task, for most cases, we want to utilize partial sequence as observation. Then it always needs to compute distances between a long sequence (from training set in a lazy learning style) and a short sequence (from the beginning part of the sample to be predicted), so local alignment based distance [48] is preferred.
- Discrete **hidden Markov model** (D-HMM) represents generative model based methods.

Fig. 12a and Table 4 shows comparison of prediction accuracy of different methods. We can clearly see that the context information, such as interactive objects, can perform as a strong cue for predicting activity type.

In HMM, high order dependencies between actionlets are lost. Though it still can get satisfactory classification performance at full observation, in terms of prediction, it has obvious limitations. In feature-based and distance-based approaches, the sequential nature of human activity cannot be captured. The context information cannot be added either. Therefore, without causality (sequential patterns) and context (object cues), things become unpredictable. Because in MPII-Cooking data set, many actionlets, such as *cut* and *take*, are actually routines in preparing different dishes, ignoring sequential patterns may result in confusion between activity classes.



(a) Activity prediction on MPII-Cooking dataset.

(b) Activity prediction on UCI-OPPORTUNITY dataset.

Fig. 12. Performance comparisons on context-aware prediction evaluation.

4.2.2 Experiments on Sensor Data Set

In this experiment, we aim to test the ability of our model to leverage from visual domain to sensor networking environment. Since this data set has no pre-segmented activity video clips, we first locate samples of each kind of high-level activity based on their annotations, such as starting and ending time stamps. Then we extract actionlets and objects labels from sensor data within the time interval of each sample. For mapping-based context-aware model, we set $\xi_{\text{itemset}} = 0.2$ and maximum PST order as 7. For SPM-based context-aware model, we set $\xi_{\text{itemset}} = 0.2$ and $\xi_{\text{sequence}} = 0.2$. We also compare our approaches with other three methods mentioned above.

Fig. 12b and Table 4 shows comparison of prediction accuracy of different methods. From the figure, we can see that the sensor data set is relatively "easier" to be predicted than the cooking activity data set. This is because the sensor data set detects actionlets and objects based on sensors, which generate less noise and can mitigate occlusions.

5 Conclusion and Future Work

In this paper, we propose a novel approach to model complex temporal composition of actionlets for activity prediction. The major contributions include a probabilistic suffix tree for representing various order Markov dependencies between action units; context-cue, especially interactive objects information, which is modeled through sequential pattern mining; and a predictive accumulative function learned from data to characterize the predictability of each kind of activity. We have empirically shown that

[6] Notice that there are many situations that some periodical actions will be segmented to consecutive duplicate eventlets, e.g., action "cut".

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

TABLE 4
Performance Comparisons on Two Context-Aware Data Sets

Methods	Cooking Activity dataset					Sensor dataset				
	20%	40%	60%	80%	100%	20%	40%	60%	80%	100%
Feature-based Model (K-grams) [14]	0.48	0.42	0.53	0.56	0.55	0.52	0.55	0.69	0.77	0.79
Distance Function base Model [48]	0.5	0.49	0.51	0.55	0.54	0.55	0.59	0.63	0.72	0.76
HMM	0.49	0.53	0.60	0.62	0.65	0.49	0.54	0.59	0.68	0.77
Action-only Model	0.58	0.59	0.64	0.64	0.66	0.62	0.65	0.75	0.77	0.78
Mapping-based Context-aware Model	**0.67**	**0.71**	**0.78**	**0.77**	**0.79**	**0.68**	**0.81**	**0.85**	**0.87**	**0.92**
SPM-based Context-aware Model	**0.69**	**0.77**	**0.79**	**0.86**	**0.88**	**0.72**	**0.84**	**0.85**	**0.92**	**0.95**

Random guess is 0.5 (Percentage as observation ratios)

incorporating causality, context-cue and predictability is particularly beneficial for predicting various kinds of human activity in diverse environments. Our approach is useful for activities with deep hierarchical structure or repetitive structure. The activities with shallow structure are not suited for this model. Also, our approach relies on a good temporal decomposition and quantization of complex activity, future work will extend this model to tolerate activities composed of noisy actionlet sequences.

ACKNOWLEDGMENTS

This research was supported in part by the NSF CNS award 1314484, Office of Naval Research award N00014-12-1-1028, Air Force Office of Scientic Research award FA9550-12-1-0201, U.S. Army Research Office grant W911NF-13-1-0160, and IC Postdoc Program Grant 2011-11071400006. The authors would thank Dr. Yu Cao and Dr. M. S. Ryoo for sharing the codes and giving suggestions to this work.

REFERENCES

[1] P. Abbeel and A.Y. Ng, "Apprenticeship Learning via Inverse Reinforcement Learning," *Proc. ACM Int'l Conf. Machine Learning*, p. 1, 2004.

[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns," *Proc. IEEE Int'l Conf. Data Eng.*, pp. 3-14, 1995.

[3] R. Agrawal et al., "Fast Algorithms for Mining Association Rules," *Proc. Int'l Conf. Very Large Data Bases*, pp. 487-499, 1994.

[4] C.L. Baker, R. Saxe, and J.B. Tenenbaum, "Action Understanding as Inverse Learning.," *J. Cognition*, vol. 113, no. 3, pp. 329-349, 2009.

[5] R. Begleiter, R. El-Yaniv, and G. Yona, "On Prediction Using Variable Order Markov Models," *J. Artificial Intelligence Research*, vol. 22, pp. 385-421, 2004.

[6] W. Brendel, A. Fern, and S. Todorovic, "Probabilistic Event Logic for Interval-Based Event Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3329-3336, 2011.

[7] W. Brendel and S. Todorovic, "Learning Spatiotemporal Graphs of Human Activities," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 778-785, 2011.

[8] P.F. Brown et al., "Class-Based N-Gram Models of Natural Language," *J. Computational Linguistics*, vol. 18, no. 4, pp. 467-479, 1992.

[9] Y. Cao et al., "Recognizing Human Activities from Partially Observed Videos," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[10] W. Choi, K. Shahid, and S. Savarese, "Learning Context for Collective Activity Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3273-3280, 2011.

[11] R. Collins, X. Zhou, and S. Teh, "An Open Source Tracking Testbed and Evaluation Web Site," *Proc. IEEE Int'l Workshop Performance Evaluation of Tracking and Surveillance*, 2005.

[12] J.W. Davis and A. Tyagi, "Minimal-Latency Human Action Recognition Using Reliable-Inference," *J. Image and Vision Computing*, vol. 24, no. 5, pp. 455-472, 2006.

[13] F. Desobry et al., "An Online Kernel Change Detection Algorithm," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2961-2974, Aug. 2005.

[14] G. Dong, *Sequence Data Mining*. Springer-Verlag, 2009.

[15] F. Facca and P. Lanzi, "Mining Interesting Knowledge from Weblogs: A Survey," *J. Data & Knowledge Eng.*, vol. 53, no. 3, pp. 225-241, 2005.

[16] Q. Fan et al., "Recognition of Repetitive Sequential Human Activity," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 943-950, 2009.

[17] A. Gaidon, Z. Harchaoui, and C. Schmid, "Actom Sequence Models for Efficient Action Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3201-3208, 2011.

[18] A. Gupta, A. Kembhavi, and L.S. Davis, "Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775-1789, Oct. 2009.

[19] P. Haider, U. Brefeld, and T. Scheffer, "Supervised Clustering of Streaming Data for Email Batch Detection," *Proc. ACM Int'l Conf. Machine Learning*, pp. 345-352, 2007.

[20] R. Hamid et al., "A Novel Sequence Representation for Unsupervised Analysis of Human Activities," *J. Artificial Intelligence*, vol. 173, no. 14, pp. 1221-1244, 2009.

[21] D. Han et al., "Selection and Context for Action Recognition," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1933-1940, 2009.

[22] M. Hoai and F. De la Torre, "Max-Margin Early Event Detectors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2863-2870, 2012.

[23] N. Ikizler-Cinbis and S. Sclaroff, "Object, Scene and Actions: Combining Multiple Features for Human Action Recognition," *Proc. European Conf. Computer Vision*, pp. 494-507, 2010.

[24] Y.A. Ivanov and A.F. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 852-872, Aug. 2000.

[25] Y. Jiang, Z. Li, and S. Chang, "Modeling Scene and Object Contexts for Human Action Retrieval with Few Examples," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 21, no. 5, pp. 674-681, May. 2011.

[26] K.-J. Kim, "Financial Time Series Forecasting Using Support Vector Machines," *J. Neurocomputing*, vol. 55, no. 1, pp. 307-319, 2003.

[27] K.M. Kitani et al., "Activity Forecasting," *Proc. European Conf. Computer Vision*, pp. 201-214, 2012.

[28] T. Kollar and N. Roy, "Utilizing Object-Object and Object-Scene Context When Planning to Find Things," *Proc. IEEE Int'l Conf. Robotics and Automation*, pp. 2168-2173, 2009.

[29] S. Kwak, B. Han, and J.H. Han, "Scenario-Based Video Event Recognition by Constraint Flow," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3345-3352, 2011.

[30] I. Laptev, "On Space-Time Interest Points," *Int'l J. Computer Vision*, vol. 64, no. 2, pp. 107-123, 2005.

[31] S. Levine, Z. Popovic, and V. Koltun, "Nonlinear Inverse Reinforcement Learning with Gaussian Processes," *Proc. Neural Information Processing Systems*, vol. 24, pp. 1-9, 2011.

[32] K. Li, J. Hu, and Y. Fu, "Modeling Complex Temporal Composition of Actionlets for Activity Prediction," *Proc. European Conf. Computer Vision*, pp. 286-299, 2012.

[33] N. Mabroukeh and C. Ezeife, "A Taxonomy of Sequential Pattern Mining Algorithms," *ACM J. Computing Surveys*, vol. 43, no. 1, p. 3, 2010.

[34] M. Marszalek, I. Laptev, and C. Schmid, "Actions in Context," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2929-2936, 2009.

[35] D. Munoz, J. Bagnell, and M. Hebert, "Stacked Hierarchical Labeling," *Proc. European Conf. Computer Vision*, pp. 57-70, 2010.

[36] D. Munoz, J.A. Bagnell, and M. Hebert, "Co-Inference for Multi-Modal Scene Analysis," *Proc. European Conf. Computer Vision*, pp. 668-681, 2012.

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.

[37] O. Nasraoui et al., “A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites,” *IEEE Trans. Knowledge and Data Eng.*, vol. 20, no. 2, pp. 202-215, Feb. 2008.

[38] D. Neill, A. Moore, and G. Cooper, “A Bayesian Spatial Scan Statistic,” *Proc. Neural Information Processing Systems*, pp. 1003-1010, 2006.

[39] J. Niebles, C.-W. Chen, and L. Fei Fei, “Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification,” *Proc. European Conf. Computer Vision*, pp. 392-405, 2010.

[40] J.C. Niebles, H. Wang, and L. Fei Fei, “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words,” *Int'l J. Computer Vision*, vol. 79, no. 3, pp. 299-318, 2008.

[41] M. Pei, Y. Jia, and S.-C. Zhu, “Parsing Video Events with Goal Inference and Intent Prediction,” *Proc. IEEE Int'l Conf. Computer Vision*, pp. 487-494, 2011.

[42] D. Roggen et al., “Collecting Complex Activity Data Sets in Highly Rich Networked Sensor Environments,” *Proc. Int'l Conf. Networked Sensing Systems*, pp. 233-240, 2010.

[43] M. Rohrbach et al., “A Database for Fine Grained Activity Detection of Cooking Activities,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1194-1201, 2012.

[44] D. Ron, Y. Singer, and N. Tishby, “The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length,” *J. Machine Learning*, vol. 25, no. 2, pp. 117-149, 1996.

[45] M.S. Ryoo, “Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos,” *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1036-1043, 2011.

[46] M.S. Ryoo and J.K. Aggarwal, “Recognition of Composite Human Activities through Context-Free Grammar Based Representation,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1709-1718, 2006.

[47] Z. Si et al., “Unsupervised Learning of Event And-Or Grammar and Semantics from Video,” *Proc. IEEE Int'l Conf. Computer Vision*, pp. 41-48, 2011.

[48] T. Smith and M. Waterman, “Identification of Common Molecular Subsequences,” *J. Molecular Biology*, vol. 147, pp. 195-197, 1981.

[49] J. Srivastava et al., “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,” *ACM SIGKDD Explorations Newsletter*, vol. 1, no. 2, pp. 12-23, 2000.

[50] P.K. Turaga, A. Veeraraghavan, and R. Chellappa, “From Videos to Verbs: Mining Videos for Activities Using a Cascade of Dynamical Systems,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[51] H. Wang et al., “Action Recognition by Dense Trajectories,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3169-3176, 2011.

[52] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining Actionlet Ensemble for Action Recognition with Depth Cameras,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1290-1297, 2012.

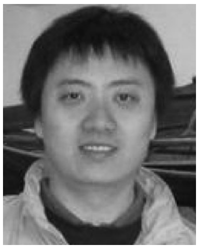
[53] K. Wang, Y. Xu, and J. Yu, “Scalable Sequential Pattern Mining for Biological Sequences,” *Proc. ACM Int'l Conf. Information and Knowledge Management*, pp. 178-187, 2004.

[54] B. Yao and L. Fei Fei, “Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 17-24, 2010.

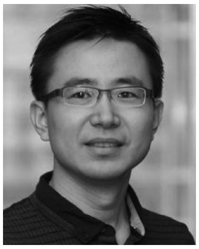
[55] Q. Zhao and S. Bhowmick, “Sequential Pattern Mining: A Survey,” Technical Report CAIS, Nayang Technological Univ., pp. 1-26, 2003.

[56] B.D. Ziebart et al., “Maximum Entropy Inverse Reinforcement Learning,” *Proc. 23rd Nat'l Conf. Artificial Intelligence-Volume 3 (AAAI '08)*, pp. 1433-1438, 2008.

[57] B.D. Ziebart et al., “Planning-Based Prediction for Pedestrians,” *Proc. IEEE Int'l Conf. Intelligent Robots and Systems*, pp. 3931-3936, 2009.



Kang Li received the BS degree in information and computational science, the MS degree in expert system and intelligent control from Northwestern Polytechnical University, China, in 2004 and 2007, respectively, and the MS degree in computer science and engineering from the State University of New York at Buffalo, Buffalo, in 2011. He is currently working toward the PhD degree in computer engineering at Northeastern University, Boston, Massachusetts. His current research interests include computer vision, applied machine learning, and data mining.



Yun Fu (S'07-M'08-SM'11) received the BEng degree in information engineering and the MEng degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, respectively, and the MS degree in statistics and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. He is an interdisciplinary faculty member affiliated with College of Engineering and the College of Computer and Information Science at Northeastern University, Boston, Massachusetts, since 2012. Prior to joining the Northeastern faculty, he was a scientist working at BBN Technologies, Cambridge, Massachussets, during 2008-2010. He held a part-time lecturer position in the Department of Computer Science, Tufts University, Medford, Massachussets, in 2009. He was a tenure-track assistant professor of the Department of Computer Science and Engineering, State University of New York, Buffalo, during 2010-2012. His research interests are interdisciplinary research in machine learning, computer vision, social media analytics, human-computer interaction, and cyber-physical systems. He is the recipient of the 2002 Rockwell Automation Master of Science Award, Edison Cups of the 2002 GE Fund Edison Cup Technology Innovation Competition, the 2003 Hewlett-Packard Silver Medal and Science Scholarship, the 2007 Chinese Government Award for Outstanding Self-Financed Students Abroad, the IEEE International Conference on Image Processing 2007 Best Paper Award, the 2007-2008 Beckman Graduate Fellowship, the 2008 M.E. Van Valkenburg Graduate Research Award, the ITESOFT Best Paper Award of 2010 IAPR International Conferences on the Frontiers of Handwriting Recognition, the 2010 Google Faculty Research Award, the IEEE International Conference on Multimedia and Expo 2011 Quality Reviewer, the IEEE ICDM 2011 Workshop on Large Scale Visual Analytics Best Paper Award, the 2011 IC Postdoctoral Research Fellowship Award, 2012 IEEE TCSVT Best Associate Editor (BAE) Award, the 2013 IEEE FG Best Student Paper Honorable Mention Award, and the 2014 INNS Young Investigator Award. He is currently an associate editor of the *IEEE Transactions on Neural Networks and Leaning Systems*, and *IEEE Transactions on Circuits and Systems for Video Technology*. He is a lifetime member of the ACM, AAAI, SPIE, and Institute of Mathematical Statistics, and a Beckman graduate fellow during 2007-2008. He is a senior member of the IEEE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.

Authorized licensed use limited to: University of Minnesota. Downloaded on August 14,2022 at 19:33:50 UTC from IEEE Xplore. Restrictions apply.