Modeling Complex Temporal Composition of Actionlets for Activity Prediction

Kang Li[1], Jie Hu[2], and Yun Fu[1]

[1] Department of ECE and College of CIS, Northeastern University, Boston, MA, USA
[2] Department of CSE, State University of New York, Buffalo, NY, USA
li.ka@husky.neu.edu, y.fu@neu.edu, jhu6@buffalo.edu

Abstract. Early prediction of ongoing activity has been more and 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. Different from early recognition on short-duration simple activities, we propose a novel framework for *long*-duration complex activity prediction by discovering the causal relationships between constituent actions and the predictable characteristics of activities. The major contributions of our work include: (1) we propose a novel activity decomposition method by monitoring motion velocity which encodes a temporal decomposition of long activities into a sequence of meaningful action units; (2) Probabilistic Suffix Tree (PST) is introduced to represent both large and small order Markov dependencies between action units; (3) we present a Predictive Accumulative Function (PAF) to depict the predictability of each kind of activity. The effectiveness of the proposed method is evaluated on two experimental scenarios: activities with middle-level complexity and activities with high-level complexity. Our method achieves promising results and can predict global activity classes and local action units.

1 Introduction

In recent years, research shows that modeling temporal structure is a basic methodology for recognition of complex human activity [1–3]. 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 sport game will be highly desirable. In public area, we want to equip a surveillance system that can raise an alarm in advance before 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 human 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

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, pp. 286–299, 2012.

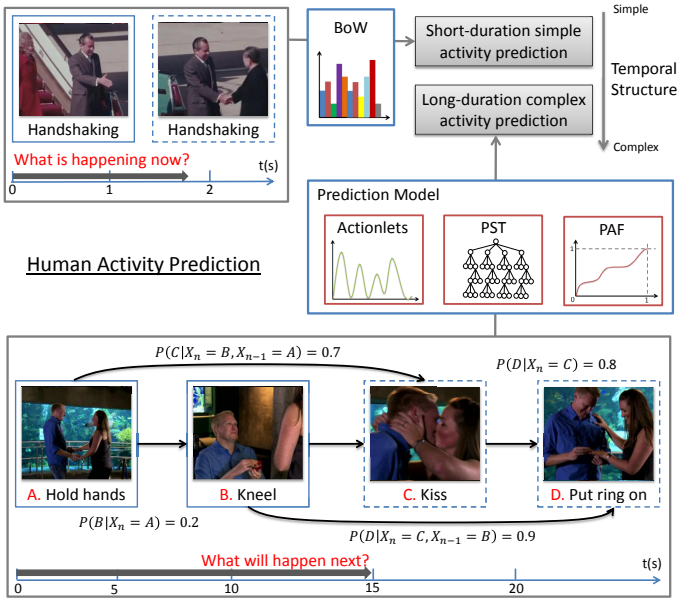


Fig. 1. Frameworks for two categories of activity prediction problems: (1) short-duration simple activity prediction (*e.g.* "handshaking"), and (2) long-duration complex activity prediction (*e.g.* "propose marriage"). The first problem can be solved in the classic bag-of-words paradigm. Our approach aims to solve the second problem.

our knowledge, the work in [4] is the only one that explicitly raised this problem. However, they identified activity prediction with only early recognition of short-duration single action, such as "hand shaking", "pushing". This assumption strongly limits the types of activity that can be predicted and the earliness of prediction. We believe that activity prediction is more desirable and valuable if it focuses on long-duration complex activities that are considered as a composition of simpler actions, such as "propose marriage".

In this paper, we propose a novel framework, shown in Fig. 1, 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 prediction with meaningful earliness and have the machine vision system provide a time-critical reaction. We represent complex activity 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 velocity. And the key contribution of this work is the idea that causality of action units can be encoded as a Probabilistic Suffix Tree (PST) with variable

temporal scale, while the predictability can be characterized by a Predictive Accumulative Function (PAF) learned from information entropy changes along every stage of activity progress. In order to test the efficacy of our method, we introduce a new dataset that focuses on complex activity in tennis game. Our method aims to answer the challenging question: "can we predict who will win?". Also we test our method on another benchmark dataset about daily indoor living activity. Our algorithm shows very promising results.

1.1 Related Work

Recently, there has been a surge in interest in complex activity recognition by involving various context 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.

The three systems that are mostly related to our work are [4, 1, 3]. [4] argues that the goal of activity prediction is to recognize unfinished activity 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. Since their model can be only suitable for prediction of instant single action, the prediction power and scalability of their approach are limited. Also they did not answer the question why a particular activity can be predicted. And the algorithm relies on the strong assumption that the distribution of local features at beginning portion of the video is discriminative for that activity.

Grammar based methods [5, 6] show effectiveness for composite human activity recognition. [3] aims 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 [7]. 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.

Syntactic model is a very powerful tool for representing activities with high level temporal logic complexity. [1] proposes the idea that global structural information of human activities can be encoded using a subset of their local event sequences. And they regarded discovering structure patterns of activity as a feature selection process. Although rich temporal structure information is encoded, they did not consider any prediction possibility from there.

Although not dealing with activity prediction directly, 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 [8, 9]. To model temporal structure of decomposable activities, [10, 11] extend the classic bag-of-words model by including segmentation and dynamic matching. [12, 13] regard complex activity recognition as a constrained optimization problem.

1.2 Dataset

Our prediction model can be applied 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 datasets with significant different temporal structure complexity. First, we collect real world video 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 intercepted 160 video clips for 160 points from a 4 hour game. Then we separate them into two categories of activity, where 80 clips are winning points and 80 clips are losing points with respect to a specific player. So our prediction problem on this dataset becomes an interesting question: "can we predict who will win?". The dataset and prediction task are illustrated in Fig. 2. Since each point consists of sequence of action units with length ranging from one to more than twenty, tennis game has a high-level temporal structure complexity in terms of both variance and order.

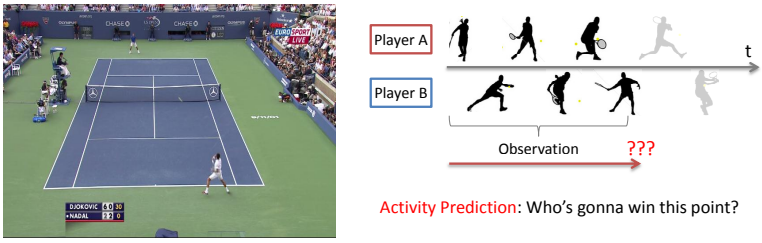


Fig. 2. Left: tennis game dataset. Right: activity prediction task on this dataset.

Second, we choose Maryland Human-Object Interactions (MHOI) dataset [14], which consists of 5 annotated activities: *answering a phone call, making phone call, drinking water, lighting a flash, pouring water into container*. These activities have about 3 to 5 action units each. And 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 9 or 10 video samples. And there are 44 video clips in total. Examples in this dataset are shown in Fig. 3.

2 Temporal Decomposition and Video Representation

2.1 Actionlets Detection by Motion Velocity

Temporal decomposition is the first key step for our representation of complex activity. It is 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 [15]. We call these atomic actions **actionlets**. We found that the

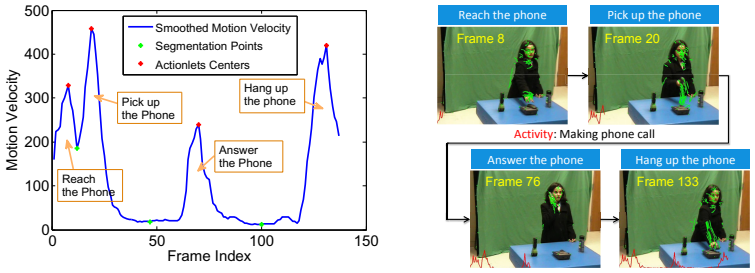


Fig. 3. Temporal decomposition of activity. We use a sample from the activity class "making a phone call" in the MHOI dataset to illustrate the proposed method. The left figure shows the smoothed motion velocity curve and corresponding detected segmentation points. The right figure shows several key frames extracted from each actionlet. The green curves around human body indicate trajectories along interest points. The red curves at the bottom of each image record the motion velocity progress.

velocity changes of human actions have similar periodic regularity. Fig. 3 shows an example of action velocity tracking and corresponding activity segmentation.

The specific method has following several steps: 1) Use Harris Corner 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:

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

where V_t represents the overall motion velocity at frame F_t, p_i is the ith interest point found in frame F_t. And $(x_{i,t}, y_{i,t})$ is the position of point p_i in the frame. We can observe 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. And the velocity reaches the peak at the intermediate stage of each action. To evaluate our temporal decomposition approach, a target window with size of 15 frames around the human labeled segmentation point would be our ground truth. We manually labeled 113 actionlets for all 44 videos in the MHOI dataset. The accuracy of automatic actionlets segmentation is **0.79**. For the tennis game dataset, we cut videos 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**.

2.2 Activity Encoding

Based on accurate temporal decomposition results, we can easily cluster actionlet into meaningful groups so that each 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 it can provide a discriminative representation for the actionlets.

Due to different spatial extent of human in the scene and different background motion styles, we use two approaches to compute descriptors for tennis game dataset and MHOI dataset respectively. For the MHOI dataset 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 HoG (Histogram of Gradients) and HoF (Histogram of Flow) descriptors [16]. Furthermore, we vector quantize the descriptors by computing memberships with respect to a descriptor codebook, which is obtained by k-means clustering of the descriptors in the training set. Then, actionlets categories are learned in an unsupervised way from histogram of spatial-temporal words [17]. To evaluate the actionlets encoding results, we manually group actionlets into 5 categories where each category represents a meaningful action unit, such as "reach the object", "grab the object", and "release the object". With a codebook size of 50, the Rand index of clustering is **0.77**. (Rand index is a measure of the similarity between data clustering and ground truth.)

For the tennis game dataset, since the scale of player in the video is very small, it is difficult to get sufficient local features by using sparse sampling methods. Here, we use dense trajectories [18] 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. Every 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 [18]. And the codebook size we used is 1000. In addition, to remove the noises caused by camera movements and shadows, a human tracker [19] is used before sampling feature records. For evaluation, we group 253 actionlets from 40 annotated videos into 10 categories, And the Rand index of clustering is **0.73**.

3 Activity Prediction 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 unsupervised learned from videos with appropriate segmentation and clustering methods. And let $D_{\text{training}} = \{r^1, r^2, ..., r^m\}$ be the training sample set of m sequences over the actionlet alphabet Σ, where the length of the ith $(i = 1, ..., m)$ sequence is l_i (*i.e.* $r^i = r_1^i r_2^i ... r_{l_i}^i, r_j^i \in \Sigma$). Based on D_{training}, the goal is to learn a model P that provides a probability assignment $p(t)$ for an ongoing actionlet sequence $t = t_1, t_2, ..., t_{\|t\|}$. To realize this design with maximum prediction power, we include two aspects 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 a satisfiable accuracy.

3.1 Pattern Knowledge Representation

Causality is an important cue for human activity prediction. So automatic acquisition of causality from sequential actionlets becomes the key. Variable order Markov Model (VMM) [20] 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 based on training data. Therefore, it can encode richer and more flexible causal relationships. Here, we model complex human activity as a Probabilistic Suffix Tree (PST) [21] which implements the single best L-bounded VMM (VMMs of degree L or less) in a fast and efficient way.

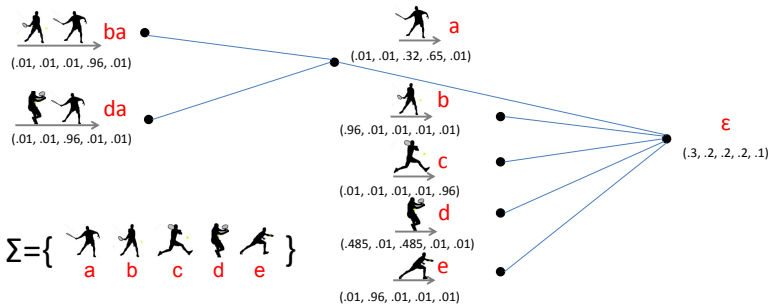


Fig. 4. An example PST corresponding to the training sequence $r = badacebadc$ over alphabet $\Sigma = \{a, b, c, d, e\}$. The vector under each node is the probability distribution over alphabet associated with the actionlets subsequence (in red). (e.g. the probability to observe d after a subsequence, whose largest suffix in the tree is ba, is 0.96).

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 next possible actionlet $\sigma \in \Sigma$. We call 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. 4 shows an example PST constructed from a training sequence of actionlets.

3.2 Predictive Accumulative Function

To characterize the predictability of activities, we formulate a Predictive Accumulative Function (PAF) in this section. We want to depict the predictable characteristic of a particular activity. For example, "tennis game" is a late-predictable problem in the sense that when we observed a long sequence of actionlets performed by two players, 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

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

1. **Forming candidate suffix set** $\overline{S}$: Let $D_{\text{training}} = \{r^1, r^2, ..., r^m\}$ be the training set, and assume s is a subsequence of r^i $(i = 1, ..., m)$. 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 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) $\frac{P(\sigma|s)}{P(\sigma|suf(s))} \geq \beta$, or $\leq 1/\beta$, which means the context s provides extra information in predicting σ relative to its longest suffix $suf(s)$. β is a user specified threshold to measure the difference between the candidate and its direct parent node.
   - **Then**, if s passed above two tests, add s and its suffixes into $\overline{T}$.
3. **Smoothing the probability distributions to get** $\gamma_{\mathbf{s}}(\sigma)$:
   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 by the user.

can guess the intention. So different activities always have quite different PAFs. In our model, PAF can be learned automatically from the training data. And 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 $r = (r_1, r_2, ..., r_l) \in D$, where $r_i (i = 1, 2, ..., l) \in \Sigma$, l is the length of r. We use $r_{\text{pre}(k)}$ 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.

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

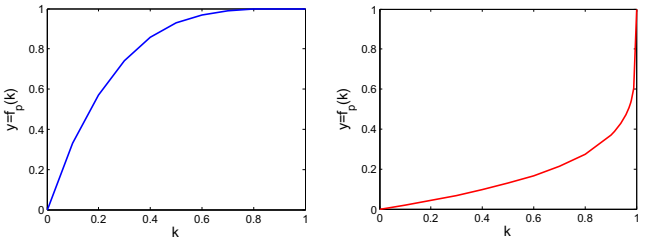
Since $r_{\text{pre}(k)}$ is the "prefix" of r, we have

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

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

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

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 the root of the tree; and $\gamma_s : \Sigma \to [0, 1]$ is the *next symbol probability function* related with s, $\sum_{\sigma \in \Sigma} \gamma_s(\sigma) = 1$.



(a) Early predictable problem (b) Late predictable problem

Fig. 5. PAFs for depicting predictable characteristics of different activities

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, we call it predictive accumulative function: $\mathbf{y} = \mathbf{f_p}(\mathbf{k})$. Function f_p depicts the predictable characteristic of a particular activity. Fig. 5 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.

3.3 Final Prediction Model

Given an ongoing sequence $t = t_1, t_2, ..., t_{\|t\|}$, we can now construct our prediction function by using the knowledge learned from Section 3.1 and Section 3.2:

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

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.

3.4 Supervised Prediction

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 $c(c = 1, ..., C)$ 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 $t = t_1, t_2, ..., t_{\|t\|}$, 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: $\mathbf{p^{\overline{T}_{c0}}(t) = Max\{p^{\overline{T}_c}(t), c = 1, ..., C\}}$.

4 Experimental Results

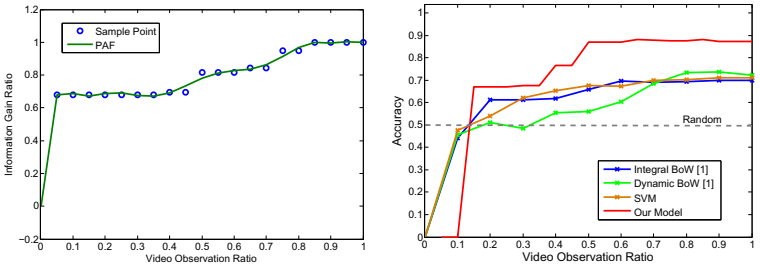
In order to test our framework, we consider two experimental scenarios. First, we test the ability of our approach to predict human activities with middle-level temporal complexity on MHOI dataset. Second, we test our model at high-level temporal complexity activities on the tennis game dataset. We show promising prediction results with our approach on both of these datasets.

4.1 Middle-Level Complex Activity Prediction

Samples in MHOI dataset are about daily activities (*e.g.* "making phone call"). This type of activity usually consists of 3 to 5 actionlets and lasts about 5 to 8 seconds, so we call it middle-level complex activity. In this dataset, each category has 9 or 10 samples. For a particular activity, we use all the samples in that category as positive set, and randomly select equal number of samples from remaining categories as negative set. Then we fit the prediction task into the context of supervised classification problem. To train a prediction model, we construct an order 5-bounded PST and fit a PAF respectively. To evaluate the prediction accuracy, we use "leave-one-out" method. Since the sample number is relatively small, we repeat our experiments 10 times for each activity and average the performance. In addition, we implemented several previous human activity prediction approaches to compare them with our method. Three types of previous prediction model using the same features were implemented: (1) Dynamic Bag-of-Words model [4], (2) Integral Bag-of-Words model [4], and (3) a basic SVM-based approach.

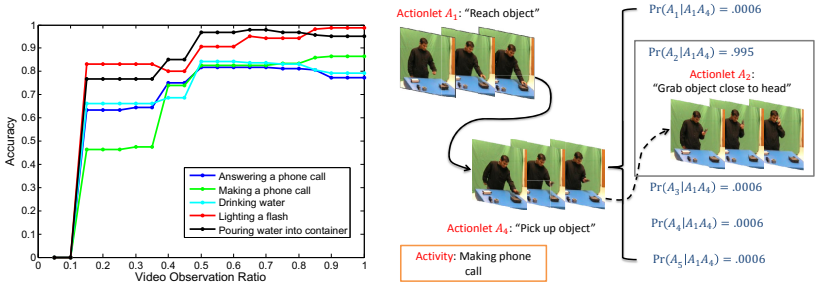
Fig. 6 (a) illustrates the process of fitting PAF from training data. It shows that daily activities such as examples from MHOI dataset are early predictable. That means the semantic information at early stage strongly exposes the intension of the whole activity. Fig. 6 (b) illustrates the performance curves of the implemented 4 methods. The results are averaged over 5 activities. Its X axis corresponds to the observed ratio of the testing videos, while the Y axis corresponds to the activity recognition accuracy of the system. The figure confirms that the proposed method has great advantages over other methods. For example, after half video observed (about 2 actionlets), our model is able to make a prediction with the accuracy of 0.9.

Fig. 7 (a) shows detailed performance of our approach over 5 different daily activities. From the figure, we can see that the activity "Pouring water into container" has the best prediction accuracy and earliness. Let's explain the reason.



(a) Fitting PAF (b) Daily activity prediction

Fig. 6. Activity prediction results on MHOI dataset. (a) shows PAF of daily activity. (b) shows comparison of prediction accuracy of different methods. A higher graph suggests that the corresponding method is able to recognize activities more accurately and earlier than the ones below it. Our approach showed the best performance by considering causal relationships among actionlets and predictable characteristics of activities.



(a) Prediction on 5 activities. (b) Predict next actionlet.

Fig. 7. Global and local prediction for a particular activity in MHOI dataset

In this dataset, after the actors reach the object, they usually then grab the object and put it close to his or her head. Three activities ("making phone call", "drinking water", and "answering phone call") share this process in the initial phase of the activity. So, in the activity "pouring water into container", after the 1st common actionlet "reach object", the 2nd and 3rd constituent actionlets make the sequence pattern quite distinctive. And Table. 1 shows prediction accuracy with respect each activity in MHOI dataset. 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. 7 (b) shows an example from our experiment results.

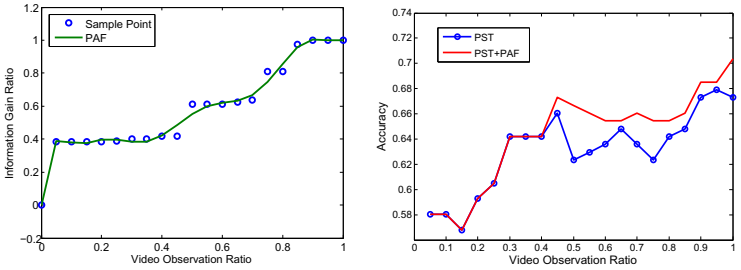
4.2 High-Level Complex Activity Prediction

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

Table 1. Performance comparisons on MHOI dataset

Methods	Answer phone		Make a call		Drinking		Lighting flash		Pouring Water	
	30% observed	50% observed	30% observed	50% observed	30% observed	50% observed	30% observed	50% observed	30% observed	50% observed
Integral BoW [1]	0.38	0.50	0.87	0.80	0.56	0.69	0.63	0.50	0.61	0.78
Dynamic BoW [1]	0.61	0.49	0.84	0.81	0.23	0.27	0.37	0.48	0.30	0.69
SVM	0.47	0.51	0.69	0.66	0.57	0.65	0.83	0.75	0.55	0.82
Our Model	**0.64**	**0.82**	**0.48**	**0.83**	**0.66**	**0.84**	**0.83**	**0.91**	**0.77**	**0.97**

dataset 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 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 6-bounded PST and a PAF are trained from data to construct the prediction model. And the same "leave-one-out" method is used for evaluation.



(a) Fitting PAF (b) Tennis game prediction

Fig. 8. Activity prediction results on tennis game dataset. (a) shows PAF of tennis game. (b) shows prediction performance of our model. We did not show comparisons with the other 3 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.

Fig. 8 (a) 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 consists with our common sense about tennis games. Fig. 8 (b) shows prediction performance of our method. Here we compare two versions of our model to illustrate the improvement caused by considering predictable characteristic of activity. Since all other three methods, D-BoW, I-BoW and SVM, failed in prediction on this dataset, we did not show comparisons. In another word, our model is the only one that has the capability to predict on high-level complex activity. Table. 2 shows detailed comparison of 4 methods on two datasets.

Table 2. Performance comparisons on two datasets. Random guess is 0.5, therefore the other 3 methods actually perform random guess on tennis game.

Methods	Tennis Game Dataset					MHOI dataset				
	20% observed	40% observed	60% observed	80% observed	100% observed	20% observed	40% observed	60% observed	80% observed	100% observed
Integral BoW [1]	0.47	0.44	0.53	0.47	0.51	0.61	0.62	0.70	0.69	0.70
Dynamic BoW [1]	0.53	0.55	0.49	0.44	0.48	0.51	0.56	0.60	0.73	0.72
SVM	0.56	0.52	0.51	0.48	0.49	0.54	0.65	0.67	0.70	0.71
Our Model	**0.59**	**0.64**	**0.65**	**0.65**	**0.70**	**0.67**	**0.77**	**0.87**	**0.88**	**0.87**

4.3 Discussions

In this subsection, we would like to discuss and highlight some aspects of our proposed activity prediction model:

1. Our approach is a general framework for activity prediction. It can be integrated with any sequential decomposition methods of complexity activity with flexible actionlets granularity.
2. Our approach is a brand new method customized to the prediction problem. Since activity classification and activity prediction are quite different problems, it is inappropriate to adopt similar bag-of-words paradigm.
3. All the experimental results validate the advantages of utilizing causality and predictability as prediction driving force, which inspires us to follow this philosophy principal when design new activity prediction techniques.
4. Compared with the only existing work [4] for activity prediction, our approach outperforms their method by a large margin on both accuracy and earliness. The proposed model is the only one that can predict on high-level complex activities.

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(PST) for representing various order Markov dependencies between action units; and a Predictive Accumulative Function(PAF) learned from data to characterize the predictability of each kind of activity. We have empirically shown that incorporating causality and predictability is particularly beneficial for predicting both middle-level complex activities as well as high-level complex activities. Our approach is useful for activities with deep hierarchical structure or repetitive structure. The activities with shallow structure are not suitable for this model. Also, our approach relies on a good temporal decomposition and quantization of complex activity, systems with a lot of errors and uncertainty are not welcome. Future directions include combining more contextual information for activity prediction, such as scene and object.

Acknowledgement. This research is supported in part by the NSF CNS 1135660, Office of Naval Research award N00014-12-1-0125, U.S. Army Research Office grant W911NF-11-1-0365, Air Force Office of Scientific Research award FA9550-12-1-0201, and IC Postdoctoral Research Fellowship award 2011-11071400006.

References

1. Hamid, R., Maddi, S., Johnson, A., Bobick, A., Essa, I., Isbell, C.: A novel sequence representation for unsupervised analysis of human activities. Artificial Intelligence 173, 1221–1244 (2009)
2. Brendel, W., Todorovic, S.: Learning spatiotemporal graphs of human activities. In: IEEE ICCV, pp. 778–785 (2011)
3. Pei, M., Jia, Y., Zhu, S.-C.: Parsing video events with goal inference and intent prediction. In: IEEE ICCV (2011)
4. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: IEEE ICCV (2011)
5. Ivanov, Y.A., Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing. IEEE PAMI 22(8), 852–872 (2000)
6. Ryoo, M.S., Aggarwal, J.K.: Recognition of composite human activities through context-free grammar based representation. In: CVPR (2006)
7. Si, Z., Pei, M., Yao, B., Zhu, S.-C.: Unsupervised learning of event and-or grammar and semantics from video. In: IEEE ICCV (2011)
8. Morariu, V.I., Davis, L.S.: Multi-agent event recognition in structured scenarios. In: CVPR (2011)
9. Brendel, W., Fern, A., Todorovic, S.: Probabilistic event logic for interval-based event recognition. In: CVPR (2011)
10. Gaidon, A., Harchaoui, Z., Schmid, C.: Actom sequence models for efficient action detection. In: CVPR (2011)
11. Niebles, J.C., Chen, C.-W., Fei-Fei, L.: Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 392–405. Springer, Heidelberg (2010)
12. Kwak, S., Han, B., Han, J.H.: Scenario-based video event recognition by constraint flow. In: CVPR (2011)
13. Fan, Q., Bobbitt, R., Zhai, Y., Yanagawa, A., Pankanti, S., Hampapur, A.: Recognition of repetitive sequential human activity. In: CVPR (2009)
14. Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: Using spatial and functional compatibility for recognition. PAMI 31, 1775–1789 (2009)
15. Turaga, P.K., Veeraraghavan, A., Chellappa, R.: From videos to verbs: Mining videos for activities using a cascade of dynamical systems. In: CVPR (2007)
16. Laptev, I.: On space-time interest points. Int. J. Comput. Vision 64, 107–123 (2005)
17. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. IJCV 79, 299–318 (2008)
18. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR (2011)
19. Collins, R., Zhou, X., Teh, S.K.: An open source tracking testbed and evaluation web site. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (2005)
20. Begleiter, R., El-Yaniv, R., Yona, G.: On prediction using variable order markov models. J. Artif. Intell. Res (JAIR) 22, 385–421 (2004)
21. Ron, D., Singer, Y., Tishby, N.: The power of amnesia: Learning probabilistic automata with variable memory length. Machine Learning 25, 117–149 (1996)