IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, Nov. 2011.

Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos

M. S. Ryoo
Electronics and Telecommunications Research Institute, Daejeon, Korea
mryoo@etri.re.kr

Abstract

In this paper, we present a novel approach of human activity prediction. *Human activity prediction is a probabilistic process of inferring* ongoing activities *from videos only containing onsets (i.e. the beginning part) of the activities. The goal is to enable early recognition of unfinished activities as opposed to the after-the-fact classification of completed activities. Activity prediction methodologies are particularly necessary for surveillance systems which are required to prevent crimes and dangerous activities from occurring. We probabilistically formulate the activity prediction problem, and introduce new methodologies designed for the prediction. We represent an activity as an integral histogram of spatio-temporal features, efficiently modeling how feature distributions change over time. The new recognition methodology named* dynamic bag-of-words *is developed, which considers sequential nature of human activities while maintaining advantages of the bag-of-words to handle noisy observations. Our experiments confirm that our approach reliably recognizes ongoing activities from streaming videos with a high accuracy.*

1. Introduction

Human activity recognition, an automated detection of events performed by humans from video data, is an important computer vision problem. In the past 10 years, the field of human activity recognition has grown dramatically, corresponding to societal demands to construct various important applications including smart surveillance, quality-of-life devices for elderly people, and human-computer interfaces. Researchers are now graduating from recognizing simple human actions such as walking and running [16, 4, 2, 10, 6], and the field is gradually moving towards recognition of complex realistic human activities involving multiple persons and objects [12, 17, 18].

Particularly, in the past 5 years, approaches utilizing spatio-temporal features obtained successful results on activity recognition in real-world environments [16, 4, 10, 6, 12, 17]. Motivated by the success of scale-invariant local patch features in object recognition, these approaches extract sparse local features from the 3-D XYT video volume formed by concatenating image frames along time axis. The bag-of-words paradigm that ignores locations of features has been widely adopted by many approaches, successfully classifying actions (e.g. videos in [16, 2]).

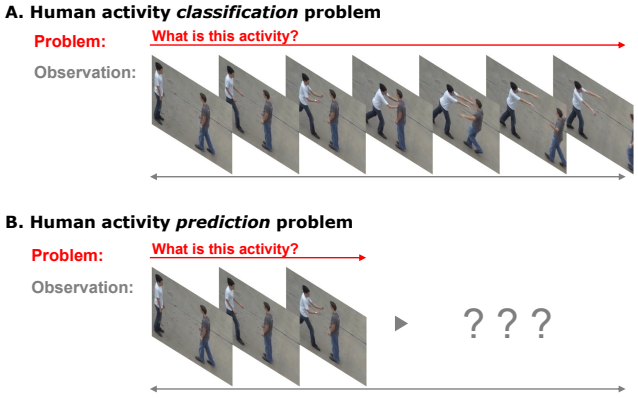

Figure 1. A comparison between the activity classification problem and the activity prediction problem. In contrast to the classification task, a system is required to infer the ongoing activity before fully observing the activity video in the prediction task.

However, most of the existing activity recognition approaches are missing an important aspect of human activity analysis. Most previous researchers including those discussed focused only on the after-the-fact detection of human activities (i.e. classifying activities after fully observing the entire video sequence), making the approaches unsuitable for the early detection of unfinished activities from video streams. In many real-world scenarios, the system is required to identify an intended activity of humans (e.g. criminals) before they fully execute the activity. For example, in a surveillance scenario, recognizing the fact that certain objects are missing after they have been stolen may not be meaningful. The system could be more useful if it is able to prevent the theft and catch the thieves by *predict-*

ing the ongoing stealing activity as early as possible based on live video observations. Similarly, if an autonomous vehicle wants to avoid an accident, its vision system is required to predict the accident which is about to occur and escape from it before any damage is caused. Even though one may extend traditional sequential models such as hidden Markov models (HMMs) to roughly approximate the prediction problem, they are unsuitable for modern high-dimensional features which provide a sparse discontinuous representation of the video. A development of a new activity prediction methodology which is able to recognize an ongoing (i.e. unfinished) activity from a video only containing early part of the activity (i.e. onsets) is necessary.

In this paper, we provide a formal definition of the *activity prediction* as an inference of the ongoing activity given temporally incomplete observations (Figure 1). The focus of this paper is the introduction of the paradigm of the activity prediction and the presentation of new methodologies designed for the prediction. Our objective is to enable the construction of an intelligent system which will perform early recognition from live video streams in real-time. We formulate the prediction problem probabilistically, and discuss the novel methodologies to solve the problem by estimating the activity's ongoing status efficiently.

This paper introduces two new human activity prediction methodologies which are able to cope with videos from unfinished activities. These methods compute the posterior probability of 'which activity has progress to which point of the activity', based on the observations available at the time. 3-D XYT spatio-temporal features strong to noise, changing background, and scale changes are adopted. We designed the activity prediction approach called *integral bag-of-words*, modeling how feature distributions of activities change as observations increase. Integral histogram representations of the activities are constructed from training videos. In addition, the new recognition methodology named *dynamic bag-of-words* approach is developed, extending the prediction algorithm to consider the sequential structure formed by video features. Structural similarities between activity models and incomplete observations are computed using our dynamic programming algorithm.

2. Previous works

Many researchers have studied human activity recognition since early 1990s [1].

As discussed in the introduction, approaches utilizing local spatio-temporal features have been popularly studied in the last 5 years [16, 4, 10, 6, 12, 17]. These features are shown to be invariant to affine transformations and robust to lighting changes, making the approaches following the bag-of-words paradigm strong to noise and changing environments. The approaches were designed to perform after-the-fact classification of activities, assuming that each video being tested contains a complete execution of a single known activity. There also have been previous works attempted to recognize activities based on video segments (e.g. a single frame [9]), which may make decisions before fully observing the activities. However, even though they obtained successful results on recognizing simple actions, they were limited in recognizing more complex activities (e.g. interactions in [13]) composed of similar body gestures.

On the other hands, previous recognition approaches using sequential state models (e.g. HMMs) [3, 8] displayed an ability to infer the intermediate status of human actions. For example, [3] modeled each human action as a sequence of hidden states generating posture features per frame, in order to enable early recognition of actions. The limitation of the previous sequential approaches is that they were unsuitable for the prediction of high-level activities with noisy observations and concurrent movements. By their nature, HMMs relied on per-frame body features of a human, and thus had difficulties processing realistic videos with changing backgrounds, dynamic lighting conditions, multiple actors, and/or unrelated pedestrians.

Recognition approaches using hierarchically organized models (e.g. stochastic context-free grammars [5]) were able to process high-level human activities by recognizing their sub-events. Particularly, Ryoo and Aggarwal's system representing human activities in terms of logical predicates [15] was able to analyze the progress status of activities based on the sub-event detection results. However, they were unable to make an appropriate analysis if human activities share similar sub-events (e.g. pointing vs. punching). Recognizing complex activities at their early stage was difficult for the previous approaches.

An important contribution of this paper is the systematic formulation of the concept of *activity prediction*, which has not been studied in depth in previous research. This paper presents novel methodologies that reliably identify unfinished activities from video streams by analyzing their onsets. Our experiments confirm that our approach is able to correctly predict ongoing activities even when the videos containing less than the first half of the activity is provided, in contrast to the previous systems.

3. Problem formulation

In this section, we probabilistically formulate the activity prediction problem. We first present our probabilistic interpretation of previous human activity classification problem briefly. Next, we formulate the new problem of human activity prediction, while contrasting it with the previous activity classification problem.

3.1. Human activity classification

The goal of human activity classification is to categorize the given videos (i.e. testing videos) into a limited number

of classes. Given a video observation O composed of image frames from time 0 to t, the system is required to select the activity label A_p which the system believes to be contained in the video. Various classifiers including K nearest neighbors (K-NNs) and support vector machines (SVMs) have been popularly used in previous approaches. In addition, sliding windows techniques were often adopted to apply activity classification algorithms for the localization of activities from continuous video streams.

Probabilistically, the activity classification is defined as a computation of the posterior probability of the activity A_p given a video O with length t. In most cases, the video duration t is ignored, assuming it is independent to the activity:

$$\begin{aligned} P(A_p \mid O, t) &= P(A_p, d^* \mid O) \\ &= \frac{P(O \mid A_p, d^*)P(A_p, d^*)}{\sum_i P(O \mid A_i, d^*)P(A_i, d^*)} \end{aligned} \tag{1}$$

where d^* is a variable describing the progress level of the activity, which indicates that the activity is fully progressed. As a result, the activity class with the maximum $P(A_p, d^* \mid O)$ is chosen to be the activity of the video O.

The probabilistic formulation of activity classification implies the classification problem assumes that each video (either a training video or a testing video) provided to the system contains a full execution of a single activity. That is, it assumes the after-the-fact categorization of video observations rather than analyzing ongoing activities, and there have been very few attempts to recognize unfinished activities.

3.2. Human activity prediction

The problem of human activity prediction is defined as an inference of unfinished activities given temporally incomplete videos. In contrast to the activity classification, the system is required to make a decision on 'which activity is occurring' in the middle of the activity execution. In activity prediction, there is no assumption that the ongoing activity has been fully executed. The prediction methodologies must automatically estimate each activity's progress status that seems most probable based on the video observations, and decide which activity is most likely to be occurring at the time. Most of the previous classification methods are not directly applicable for this purpose, since they only support the after-the-fact detections.

We probabilistically formulate the activity prediction process as:

$$\begin{aligned} P(A_p \mid O, t) &= \sum_d P(A_p, d \mid O, t) \\ &= \frac{\sum_d P(O \mid A_p, d)P(t \mid d)P(A_p, d)}{\sum_i \sum_d P(O \mid A_i, d)P(t \mid d)P(A_i, d)} \end{aligned} \tag{2}$$

Figure 2. Example 3-D spatio-temporal features (left) and visual words (right) from a hand-shake video. Features grouped into an identical visual word are described with the same color.

where d is a variable describing the progress level of the activity A_p. For example, $d = 50$ indicates that the activity A_p has been progressed from the 0th frame to the 50th frame of its representation. That is, it describes that the activity prediction process must consider various possible progress statuses of the activities for all $0 \leq d \leq d^*$. $P(t|d)$ represents the similarity between the length of the observation t and that of the activity progress d.

The key of the activity prediction problem is the accurate and efficient computation of the likelihood value $P(O|A_p, d)$, which measures the similarity between the video observation and the activity A_p having the progress level of d. A brute force method of solving the activity prediction problem is to construct multiple probabilistic classifiers (e.g. probabilistic SVMs) for all possible values of A_p and d. However, training and maintaining hundreds of classifiers to cover all progress level d (e.g. 300 SVMs per activity if the activity takes 10 seconds in 30 fps) requires a significant amount of computational costs. Furthermore, the brute force construction of independent classifiers ignores sequential relations among the likelihood values, making the development of robust and efficient activity prediction methodologies necessary.

4. Prediction using integral bag-of-words

In this section, we present our human activity prediction methodology named *integral bag-of-words*. The major difference between the approach introduced in this section and the previous approaches is that our approach is designed to efficiently analyze the status of ongoing activities from video streams. In Subsection 4.1, we first discuss the video features used. Next, our new probabilistic activity prediction methodology is presented in Subsection 4.2.

4.1. Features and visual words

Our approach takes advantage of 3-D space-time local features to predict human activities. A spatio-temporal feature extractor (e.g. [16, 4]) detects interest points with salient motion changes from a video, providing descriptors that represent local movements occurring in a video. The feature extractor first converts a video into the 3-D XYT

volume formed by concatenating image frames along time axis, and locates 3-D volume patches with salient motion changes. A descriptor is computed for each local patch by summarizing gradients inside the patch.

Once local features are extracted, our method clusters them into multiple representative types based on their appearance (i.e. feature vector values). These types are called 'visual words', which essentially are clusters of features. We use k-means clustering algorithm to form visual words from features extracted from sample videos. As a result, each detected feature in a video belongs to one of the k visual words. Figure 2 shows example features and words.

4.2. Integral bag-of-words

Integral bag-of-words is a probabilistic activity prediction approach that constructs integral histograms to represent human activities. In order to predict the ongoing activity given a video observation O of length t, the system is required to compute the likelihood $P(O|A_p, d)$ for all possible progress level d of the activity A_p. What we present in this subsection is an efficient methodology to compute the activity likelihoods by modeling each activity as an integral histogram of visual words.

Our integral bag-of-words method is a histogram-based approach, which probabilistically infers ongoing activities by computing the likelihood $P(O|A_p, d)$ based on feature histograms. The idea is to measure the similarity between the video O and the activity model (A_p, d) by comparing their histogram representations. The advantage of the histogram representation is that it is able to handle noisy observations with varying scales. For all possible (A_p, d), our approach computes the histogram of the activity, and compares them with the histogram of the testing video.

A feature histogram is a set of k histogram bins, where k is the number of visual words (i.e. feature types). Given an observation video, each histogram bin counts the number of extracted features with the same type, ignoring their spatio-temporal locations. The histogram representation of an activity model (A_p, d) is computed by averaging the feature histograms of training videos while discarding features observed after the time frame d. That is, each histogram bin of the activity model (A_p, d) describes the expected number of corresponding visual word's occurrences, which will be observed if the activity A_p has progress to the frame d.

In order to enable the efficient computation of likelihoods for any (A_p, d) using histograms, we model each activity by constructing its *integral histogram*. Formally, an integral histogram of a video is defined as a sequence of feature histograms, $H(O_l) = [h_1(O_l), h_2(O_l), ..., h_{|H|}(O_l)]$, where $|H|$ is the number of frames of the activity video O_l. Let v_w denote the wth visual word. Then, a value of the wth histogram bin of each histogram $h_d(O_l)$ is computed as:

$$h_d(O_l)[w] = |\{f \mid f \in v_w \wedge t_f < d\}| \tag{3}$$

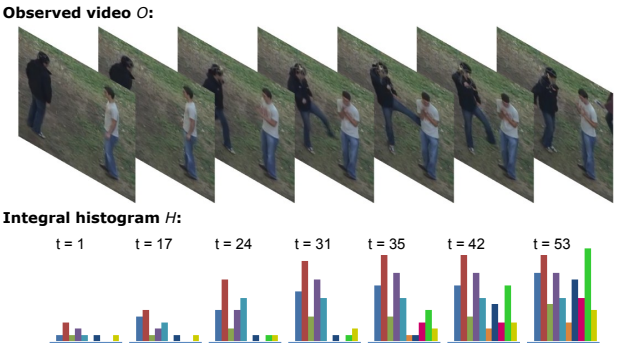

Figure 3. An example integral histogram representing a kicking video. An integral histogram models how histogram distribution changes over time. Each histogram bin counts the number of features grouped into its visual word.

where f is a feature extracted from the video O_l and t_f is its temporal location. That is, each element $h_d(O_l)$ of the integral histogram $H(O_l)$ describes the histogram distribution of spatio-temporal features whose temporal locations are less than d. Our integral histogram can be viewed as a temporal version of the spatial integral histogram [11].

Figure 3 shows an example integral histogram. Essentially, an integral histogram is a function of time describing how histogram values change as the observation duration increases. The integral histograms are computed for all training videos of the activity, and their mean integral histogram is used as a representation of the activity. The idea is to keep track of changes in the visual words being observed as the activity progress.

The constructed integral histograms enable us the prediction of human activities. Modeling integral histograms of activities with Gaussian distributions having a uniform variance, the problem of predicting the most probable activity A^* is enumerated from Equation (2) as follows:

$$\begin{aligned} A^* &= \arg\max_p \sum_d P(A_p, d \mid O, t) \\ &= \arg\max_p \frac{\sum_d M(h_d(O), h_d(A_p))P(t \mid d)}{\sum_i \sum_d M(h_d(O), h_d(A_i))P(t \mid d)} \end{aligned} \tag{4}$$

where

$$M(h_d(O), h_d(A_i)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(h_d(O)-h_d(A_i))^2}{2\sigma^2}}. \tag{5}$$

Here, $H(A_i)$ is the integral histogram of the activity A_i, and $H(O)$ is the integral histogram of the video O. An equal prior probability among activities is assumed, and σ^2 describes the uniform variance.

The proposed method is able to compute the activity likelihood for all d with $O(k \cdot d^*)$ computations given the integral histogram of the activity. The time complexity of the integral histogram construction for each activity is

$O(m \cdot log\, m + k \cdot d^*)$ where m is the number of total features in training videos of the activity. That is, our approach requires significantly less amount of computations compared to the brute force method of applying previous classifiers. For instance, the brute force method of training SVMs for all d takes $O(n \cdot k \cdot d^* \cdot r)$ computations where n is the number of training videos of the activity and r is the number of iterations to train a SVM.

5. Prediction using dynamic bag-of-words

In this section, we present a novel activity recognition methodology, *dynamic bag-of-words*, which predicts human activities from onset videos using a sequential matching algorithm. The integral bag-of-words presented in the previous section is able to perform an activity prediction by analyzing ongoing status of activities, but it ignores temporal relations among extracted features. In Subsection 5.1, we introduce a new concept of dynamic bag-of-words that considers human activities' sequential structures for the prediction. Subsection 5.2 presents our dynamic programming implementation to predict ongoing activities from videos.

5.1. Dynamic bag-of-words

Our dynamic bag-of-words is a new activity recognition approach that considers the sequential nature of human activities, while maintaining the bag-of-words' advantages to handle noisy observation. An activity video is a sequence of images describing human postures, and its recognition must consider the sequential structure displayed by extracted spatio-temporal features. The dynamic bag-of-words method follows our prediction formulation (i.e. Equation (2)), measuring the posterior probability of the given video observation generated by the learned activity model. Its advantage is that the likelihood probability, $P(O|A_p, d)$, is now computed to consider the activities' sequential structures.

Let Δd be a sub-interval of the activity model (i.e. A_p) that ends with d, and let Δt be a sub-interval of the observed video (i.e. O) that ends with t. In addition, let us denote the observed video O more specifically as O^t, indicating that O is obtained from frames 0 to t. Then, the likelihood between the activity model and the observed video can be enumerated as:

$$P(O^t \mid A_p, d) = \sum_{\Delta t} \sum_{\Delta d} [P(O^{t-\Delta t} \mid A_p, d - \Delta d) \; P(O^{\Delta t} \mid A_p, \Delta d)] \tag{6}$$

where $O^{\Delta t}$ corresponds to the observations obtained during the time interval of Δt, and $O^{t-\Delta t}$ corresponds to those obtained during the interval $t - \Delta t$.

The idea is to take advantage of the likelihood computed for the previous observations (i.e. $P(O^{t-\Delta t} \mid A_p, d - \Delta d)$) to update the likelihood of the entire observations (i.e. $P(O^t|A_p, d)$). This incremental likelihood computation not only enables efficient activity prediction for increasing observations, but also poses a temporal constraint that observations must match the activity model sequentially.

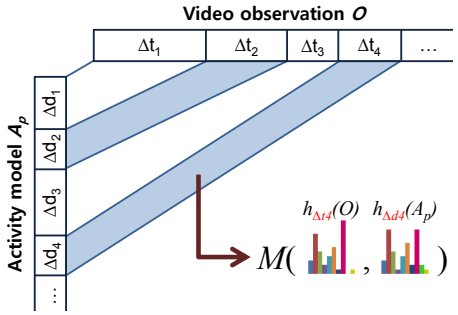

Figure 4. The matching process of our dynamic bag-of-words.

Essentially, the above-mentioned recursive equation is dividing the activity progress time interval d into a set of sub-intervals $D = \{\Delta d_1, \Delta d_2, ..., \Delta d_q\}$ with varying lengths and the observed video O into a set of sub-intervals $T = \{\Delta t_1, \Delta t_2, ..., \Delta t_q\}$. The likelihood is being computed by matching the q pairs of sub-intervals $(\Delta d_j, \Delta t_j)$. That is, the method searches for the optimal D and T that maximize the overall likelihood between two sequences, which is measured by computing similarity between each $(\Delta d_j, \Delta t_j)$ pair. Figure 4 illustrates such process.

The motivation is to divide the activity model and the observed sequence into multiple segments to find the structural similarity between them. Notice that the duration of the activity model segment (i.e. Δd) to match the new observation segment (i.e. $O^{\Delta t}$) is dynamically selected, finding the best-matching segment pairs to compute their similarity distance recursively. The segment likelihood, $P(O^{\Delta t}|A_p, \Delta d)$, is computed by comparing their histogram representations. That is, the bag-of-words paradigm is applied for matching the interval segments, while the segments themselves are sequentially organized based on our recursive activity prediction formulation.

Video segment matching using integral histograms. Our dynamic bag-of-words method takes advantage of integral histograms for computing the similarity between interval segments (i.e. $P(O^{\Delta t}|A, \Delta d)$). Integral histograms enable efficient constructions of the histogram of the activity segment Δd and that of the video segment Δt for any possible $(\Delta d, \Delta t)$. Let $[a, b]$ be the time interval of Δd. The histogram corresponding to Δd is computed as:

$$h_{\Delta d}(A_p) = h_b(A_p) - h_a(A_p) \tag{7}$$

where $H(A_p)$ is the integral histogram of the activity A_p. Similarly, the histogram of Δt is computed based on the integral histogram $H(O)$, providing us $h_{\Delta t}(O)$.

Using the integral histograms, the likelihood probability computation of our dynamic bag-of-words is described with the following recursive equation. Similar to the case of integral bag-of-words method, the feature histograms of the activities are modeled with Gaussian distributions.

$$F_p(t,d) = \sum_{\Delta t}\sum_{\Delta d} [F_p(t-\Delta t, d-\Delta d)\cdot M\left(h_{\Delta t}(O), h_{\Delta d}(A_p)\right)] \quad (8)$$

where $F_p(t,d)$ is equivalent to $P(O^t \mid A_p, d)$.

5.2. Dynamic programming algorithm

In this subsection, we present a dynamic programming implementation of our dynamic bag-of-words method to find the ongoing activity from the given video. We construct the maximum a posteriori (MAP) classifier of deciding which activity is mostly likely to be occurring.

Given the observation video O with length t, the activity prediction problem of finding the most probable ongoing activity A^* is expressed as follows:

$$A^* = \arg\max_p \frac{\sum_d F_p(t,d)P(t \mid d)P(A_p,d)}{\sum_i \sum_d F_i(t,d)P(t \mid d)P(A_i,d)}. \quad (9)$$

That is, in order to predict the ongoing activity given an observation O^t, the system is required to calculate the likelihood $F_p(t,d)$ (i.e. Equation (8)) recursively for all activity progress status d.

However, even with integral histograms, brute force searching of all possible combination of $(\Delta t, \Delta d)$ for a given video of length t requires $O(k \cdot (d^*)^2 \cdot t^2)$ computations: In order to find A^* at each time step t, the system must compute $F_p(t,d)$ for d^* number of possible d. Furthermore, computation of each $F_p(t,d)$ requires the summation of F_p values of all possible combination of Δt and Δd, as we described in Equation (8).

In order to make the prediction process computationally tractable, we design an algorithm that approximates the likelihood $F_p(t,d)$ by making its Δt to have a fixed duration. The image frames of the observed video are divided into several segments with a fixed duration (e.g. 1 second), and they are matched with the activity segments dynamically. Let u be the variable describing the unit time duration. Then, the activity prediction likelihood is approximated as:

$$F'_p(u,d) = \max_{\Delta d} F'_p(u-1, d-\Delta d) M(h_{\tilde{u}}(O), h_{\Delta d}(A_p)) \quad (10)$$

where $\tilde{u}$ is a unit time interval between $u-1$ and u.

Our algorithm sequentially computes $F'_p(u,d)$ for all u. At each iteration of u, the system searches for the best-matching segment Δd for $\tilde{u}$ that maximizes the function F', as described in Equation (10). Essentially, our method interprets a video as a sequence of ordered sub-intervals (i.e. $\tilde{u}$) where each of them is represented with a histogram of features in it. As a result, $F'_p(u,d)$ provides us an efficient approximation of the activity prediction likelihood, measuring how probable the observation O is generated from the activity (i.e. A_p) progressed to the dth frame.

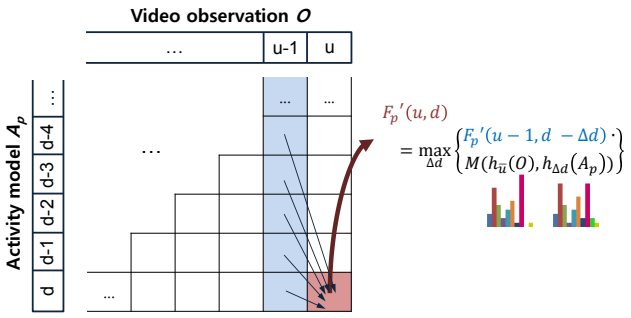

Figure 5. A figure illustrating the process of our dynamic programming algorithm. The algorithm iteratively fills in the array to obtain the optimum likelihood.

A traditional dynamic programming algorithm that corresponds to the above recursive equation is designed to calculate the likelihood. The goal is to search for the optimum activity model divisions (i.e. Δd) that best describes the observation, matching them with the observation stage-by-stage. Figure 5 illustrates the process of our dynamic programming algorithm to compute the likelihood of the ongoing activity from an incomplete video. The time complexity of the algorithm is $O(k \cdot (d^*)^2)$ for each time step u, which is in general much smaller than t.

6. Experiments

In this section, we implement and evaluate our human activity prediction methodologies while comparing them with other previous classification works. The methods' ability to perform early detection of activities is tested with the public video dataset containing high-level multi-person interactions. We confirm the advantages of our approaches on inferring ongoing activities at their early stage.

6.1. Dataset

For our experiments, we used the segmented version of the UT-Interaction dataset [13] containing videos of six types of human activities: hand-shaking, hugging, kicking, pointing, punching, and pushing. The UT-Interaction dataset is a public video dataset containing high-level human activities of multiple actors. The dataset is composed of two different sets with different environments (Figure 6), containing a total of 120 videos of six types of human-human interactions. Each set is composed of 10 sequences, and each sequence contains one execution per activity. The videos involve camera jitter and/or background movements (e.g. trees). Several pedestrians are present in the videos

Snapshot from the UT-Interaction set #1

Snapshot from the UT-Interaction set #2

Figure 6. Example snapshots from the UT-Interaction dataset.

as well, preventing the recognition. The UT-Interaction dataset was used for the human activity recognition contest (SDHA 2010) [14], and it has been tested by several state-of-the-art methods [17, 18, 19].

We chose a dataset composed of complex activities having sufficient temporal durations, instead of testing our system with videos of periodic and instantaneous actions. Even though the KTH dataset [16] and the Weizmann dataset [2] have been popularly used for the action classification in previous works, they were inappropriate for our experiments: Their videos are composed of short periodic movements which only require few frames (e.g. a single frame [9]) to perform a reliable recognition.

6.2. Experimental settings

We implemented two human activity prediction systems based on the two methods presented in this paper: the integral bag-of-words (BoW) method and the dynamic BoW method. We adopted the cuboid feature descriptors [4] as spatio-temporal features used by the systems. In principle, our proposed approaches are able to cope with any spatio-temporal feature extractors as long as they provide XYT locations of the features being detected. Extracted features were then clustered into 800 visual words (i.e. $k = 800$), and integral histograms were constructed.

In addition, we implemented several previous human activity classification approaches to compare them with our methods. Four types of previous classifiers using the same features (i.e. cuboid features) were implemented: We implemented (i) a brute-force prediction approach of constructing SVMs for all progress levels (BP-SVMs) [7] and (ii) a basic voting-based approach that casts a probabilistic vote for each spatio-temporal feature. In addition, in order to show the limitation of the previous classification framework, we implemented (iii) Bayesian classifiers with Gaussian models and (iv) standard SVMs, which are designed to classify videos assuming that they contain full activity executions. The testing was performed by applying the learned classifiers to the videos containing ongoing activities.

Throughout our experiments, the leave-one-sequence-out cross validation setting was used to measure the performances of the systems. There are 10 sequences (each sequence contains 6 videos) for each UT-Interaction dataset,

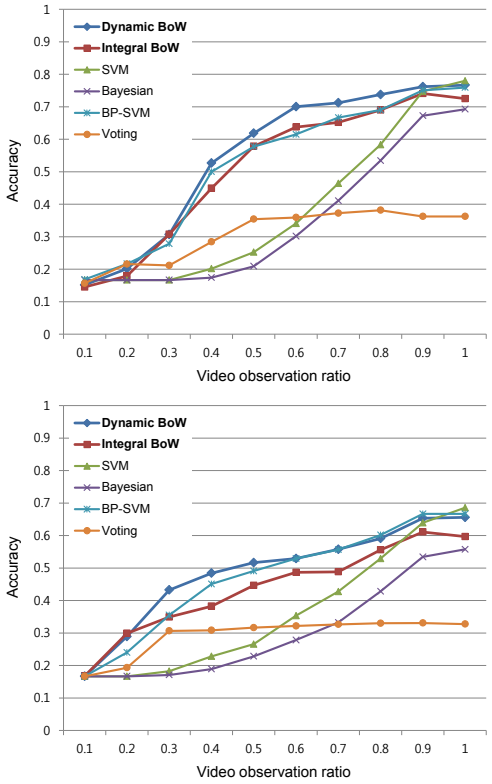

Figure 7. Activity recognition performances with respect to the observed ratio, tested with the UT-Interaction dataset #1 (top) and #2 (bottom). A higher graph suggests that the corresponding method is able to recognize activities more accurately and earlier than the ones below it. Our dynamic BoW showed the best performance by considering sequential relations among the feature points. BP-SVMs, which require a large amount of computations, also showed a fair performance. However, it generated inferior results particularly when making early decision. We are also able to observe that the per-feature-voting approach is able to recognize human activities at relatively early stage than the other basic classifiers (e.g. SVMs), but it display worse performances overall because of the characteristics of the activities in the UT-Interaction dataset (i.e. they share many gestures, generating similar features).

and thus 10-fold cross validation is performed. That is, for each round, videos in one sequence was selected as the testing videos, and videos in the other sequences were used for the training. Integral histograms were constructed from the training videos to recognize activities in the testing videos. This testing/training video selection process was repeated for 10 rounds, measuring the average recognition accuracy.

6.3. Results

In order to test the systems' ability to predict ongoing activities at their earlier stage, we measured the systems' recognition performances for classifying videos of incomplete activity executions. The experiments were conducted with 10 different observation ratio settings, from 0.1 to 1.

Let c denote the length of an original video containing a fully executed activity. Then, the setting with an observed ratio of x indicates that the video provided to the system is formulated by segmenting the initial $x \cdot c$ frames of the original video. For example, the systems' performances at observed ratio 0.5 describe the classification accuracies given testing videos only having the first halves of the activities. The observed ratio of 1 indicates that all testing videos contain full executions of the activities, making the problem a conventional activity classification problem.

Figure 7 illustrates the performance curves of the implemented systems. 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. We have averaged the systems' performances for 20 runs, considering that the visual word clustering contains randomness.

The figure confirms that the proposed methods are able to recognize ongoing activities at earlier stage of the activities, compared to the previous approaches. For example, in dataset #1, the dynamic BoW is able to make a prediction with the accuracy of 0.7 after observing only first 60% of the testing video, while the BP-SVMs must observe the first 82% to obtain the same accuracy. In addition, our integral BoW and dynamic BoW are computationally more efficient than the previous brute-force approaches (e.g. BP-SVMs). The time complexity to construct activity models in our two methods is $O(m \cdot log\, m + k \cdot d^*)$ while that of the BP-SVMs is $O(n \cdot k \cdot d^* \cdot r)$, as discussed in Subsection 4.2.

Table 1 compares the classification accuracies measured with the UT-Interaction #1. Each accuracy in the table can be viewed as the highest performance one can get using the method with the optimum parameters and visual words. Together with the classification accuracies given videos containing full activity executions, the table lists the prediction accuracies measured with the videos of the observed ratio 0.5. We are able to observe that our approaches perform superior to the previous methods given the videos with the observed ratio 0.5, confirming that they predict ongoing activities earlier. In the traditional classification task (i.e. full videos), our approach performed comparable to the state-of-the-arts results.

Our algorithms run in real-time with our unoptimized C++ implementation, except for the adopted feature extraction component. That is, if appropriate features are provided to our system in real-time, the system is able to analyze input videos and predict ongoing activities in real-time.

Table 1. Recognition performances measured with the UT-Interaction dataset #1. The classification accuracies of our approaches and the previous approaches are compared. Most of the listed classification approaches used 3-D spatio-temporal features. In addition, [17] took advantage of automatically extracted bounding boxes of persons.

System	Accuracy w. half videos	Accuracy w. full videos
Dynamic BoW	**70.0 %**	**85.0 %**
Integral BoW	**65.0 %**	**81.7 %**
Waltisberg *et al.* [17]	-	88.0 %
Cuboid + SVMs [14]	31.7 %	85.0 %
BP-SVM [7]	65.0 %	83.3 %
Yu *et al.* [18]	-	83.3 %
Yuan *et al.* [19]	-	78.2 %
Cuboid + Bayesian	25.0 %	71.7 %

7. Conclusion

In this paper, we introduced the new paradigm of human activity prediction. The motivation was to enable the early detection of unfinished activities from initial observations. We formulated the problem probabilistically, and presented two novel recognition methodologies designed for the efficient prediction of human activities. The experimental results confirmed that the proposed approaches are able to recognize ongoing human-human interactions at their much earlier stage than the previous methods.

References

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

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.

[3] J. W. Davis and A. Tyagi. Minimal-latency human action recognition using reliable-inference. *Image Vision Computing*, 24:455–472, May 2006.

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

[5] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE T PAMI*, 22(8):852–872, 2000.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[7] K. Laviers, G. Sukthankar, M. Molineaux, and D. W. Aha. Improving offensive performance through opponent modeling. In *Artificial Intelligence for Interactive Digital Entertainment Conference*, 2009.

[8] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and Viterbi path searching. In *CVPR*, 2007.

[9] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human actions classification. In *CVPR*, 2007.

[10] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3), Sep 2008.

[11] F. Porikli. Integral histogram: A fast way to extract histograms in Cartesian spaces. In *CVPR*, 2005.

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

[13] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR Contest on Semantic Description of Human Activities (SDHA), 2010.

[14] M. S. Ryoo, C.-C. Chen, J. K. Aggarwal, and A. Roy-Chowdhury. An overview of contest on semantic description of human activities (SDHA) 2010. In *Proceedings of ICPR Contests*, 2010.

[15] M. S. Ryoo, K. Grauman, and J. K. Aggarwal. A task-driven intelligent workspace system to provide guidance feedback. *CVIU*, 114(5):520–534, May 2010.

[16] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.

[17] D. Waltisberg, A. Yao, J. Gall, and L. V. Gool. Variations of a Hough-voting action recognition system. In *ICPR Contest on Semantic Sescription of Human Activities (SDHA), in Proceedings of ICPR Contests*, 2010.

[18] T.-H. Yu, T.-K. Kim, and R. Cipolla. Real-time action recognition by spatiotemporal semantic and structural forests. In *BMVC*, 2010.

[19] F. Yuan, V. Prinet, , and J. Yuan. Middle-level representation for human activities recognition: the role of spatio-temporal relationships. In *ECCV Workshop on Human Motion: Understanding, Modeling, Capture and Animation*, 2010.