Predicting Human Activities using Spatio-Temporal Structure of Interest Points

Gang Yu, Junsong Yuan
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
gyu1@e.ntu.edu.sg, jsyuan@ntu.edu.sg

Zicheng Liu
Microsoft Research
Redmond, WA, USA
zliu@microsoft.com

ABSTRACT

Early recognition and prediction of human activities are of great importance in video surveillance, e.g., by recognizing a criminal activity at its beginning stage, it is possible to avoid unfortunate outcomes. We address early activity recognition by developing a Spatial-Temporal Implicit Shape Model (STISM), which characterizes the space-time structure of the sparse local features extracted from a video. The early recognition of human activities is accomplished by pattern matching through STISM. To enable efficient and robust matching, we propose a new random forest structure, called multi-class balanced random forest, which makes a good trade-off between the balance of the trees and the discriminative abilities. The prediction is done simultaneously for multiple classes, which saves both the memory and computational cost. The experiments show that our algorithm significantly outperforms the state of the arts for the human activity prediction problem.

Categories and Subject Descriptors

H.4.0 [[**INFORMATION SYSTEMS APPLICATIONS**]: General

Keywords

Action Prediction, Random Forest, Hough Voting

1. INTRODUCTION

Surveillance cameras are widely used nowadays and the functions of these cameras are mainly for security monitoring and detecting illegal actions and events. With more and more equipped surveillance cameras, it is of great interests to develop intelligent video surveillance that can sense and understand the human activities. Although many activity recognition methods have been proposed to enable intelligent surveillance, most of them only provide after-the-fact classification of completed activities [5][10]. However, in many surveillance scenarios, it is desirable to recognize activities even before they are completed. For instance, in a supermarket, it is better to send off an alarm while someone is stealing rather than after the stealing, because it can possibly prevent this criminal activity and also provide more time for the security guard to react. As another example shown in Fig. 1, when there are people fighting on a street, it is extremely useful to recognize and stop the fighting activity early before the situation becomes worse.

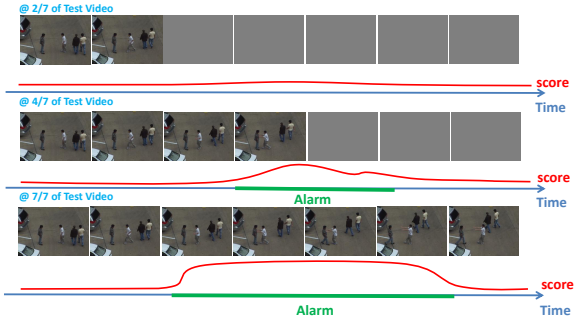

Figure 1: An illustration of the human activity prediction problem. We want to predict the "push" activity and seven sample frames are selected among the testing video. Three experiments on different observation ratios (at $2/7, 4/7, 7/7$) are shown with the red curve describing the score at each frame. The green solid line on the time coordinate refers to the predicted activity in the testing video.

The problem of human activity prediction has been proposed in [6]: *inference of the ongoing activity given temporally incomplete observations.* Integral bag-of-words (BoW) and dynamic bag-of-words are proposed in [6] to enable activity prediction with only partial observations. Despite certain successes of [6], it still has several limitations. First, since the BoW model ignores the spatial-temporal relationships among interest points, it is not discriminative enough to describe human activities. Also, although integral BoW and dynamic BoW in [6] consider the temporal information by matching between sub-intervals, there lacks a principled way to determine the optimal interval length. Finally, as we usually have a large number of categories of activities to detect, it demands an algorithm whose computational complexity is sub-linear or constant to the number of categories.

To address these limitations, we propose Spatial-Temporal Implicit Shape Model (STISM), which can well model the relationships between the local features and efficiently predict multiple activities simultaneously. To enable efficient and robust matching, a new type of random forest is proposed,

MM'12, October 29–November 2, 2012, Nara, Japan.

which makes a good trade-off between the tree balance and discriminative ability. Meanwhile, the trees will be trained for multi-class purpose, which makes our algorithm scalable to the number of classes. Given a normal desktop PC, our human activity prediction algorithm can be run in real-time. In addition to the speed benefit, STISM makes it possible to progressively predict the human activities thanks to the additive nature of the model. Even when we only have partial observation, the prediction can be accurate as well. Our action prediction experiments on UT-Interaction dataset [7] further validate the performance of our algorithm.

1.1 Activity Prediction Problem

Fig. 2 summarizes the differences among human action classification, action detection and action prediction (action and activity are used interchangeably). In the past decade, there have been a lot of great works in the human action classification [1][5][10][9][13]. The goal of human action classification is to determine the action category of a segmented video clip, referred to as X, among a set of classes $\{1, \cdots, K\}$. For the action detection [3][11], it aims to not only determine the category but also localize the position of the human action, i.e., $f^c(X) \rightarrow [x, w, y, h, t, d]$, where $f^c(X)$ refers to the detection function for the action category c and $[x, w, y, h, t, d]$ refers to the 3D position of the human action (center position x, y, t and scale w, h, d).

Action prediction [6], however, is more challenging because we need to efficiently predict the action given incomplete observations. Compared with action detection, action prediction focuses more on determining whether or not a specified action is happening. Thus, it does not require the exact localization of the human actions. In our paper, instead of determining the exact $3D$-volume of the human actions, we focus on determining the existence of an action given incomplete observations. Formally, we want to find a function:

$$f(O) \rightarrow \{0, 1, \cdots, K\}, \tag{1}$$

where $O \subset X$ refers to the incomplete observations, and $\{1, \cdots, K\}$ refers to the category of the predicted action while 0 means no target action is happening. Observation O can be either on segmented videos similar to action classification or unsegmented videos similar to action detection. In this paper, we only focus on predicting actions with segmented videos, but our algorithm can be extended to handle the case when the testing videos are not segmented.

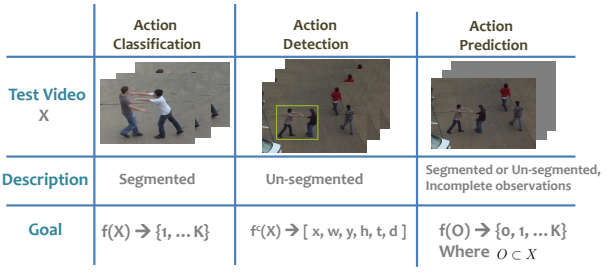

Figure 2: Comparison of action classification, action detection, and action prediction.

2. PROBLEM FORMULATION

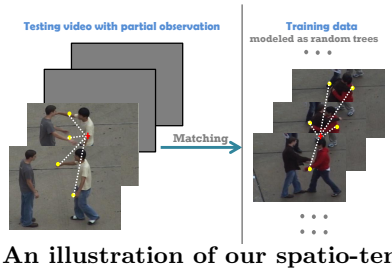

Figure 3: An illustration of our spatio-temporal activity matching.

We represent the videos with spatial-temporal interest point (STIP) [1] due to its sparsity and good performance for action recognition. Other types of local features are also applicable to our algorithm. Given a video, we refer our Spatial-Temporal Implicit Shape Model (STISM) as $\mathcal{V} = \{(f_i, s_i, c)\}$, where f_i refers to the feature description, $s_i = l_i - l_\mathcal{V}$ refers to the spatio-temporal location shift from the ith STIP position (l_i) to the center position of video $l_\mathcal{V}$, and c refers to the category of the video. STISM is a 3D extension of implicit shape model in [2]. Fig. 3 illustrates the idea of using STISM for activity matching. The yellow dots refer to the detected interest points and the white dash lines refer to the shift from the interest point to the video center, i.e., $l_i - l_\mathcal{V}$. The benefits of our implicit shape model are two-fold. On one hand, it is flexible to utilize the spatial-temporal configuration of the interest points for recognition. More specifically, we do not need to explicitly define and learn a model. On the other hand, the computational cost is low which enables real-time activity prediction. Our goal is, given a training set $\mathcal{D} = \{(f_j, s_j, c_j)\}$ (several different f_j will share the same video center location and c_j if they are from the same training video), to determine the category of c for testing video $\mathcal{V}$. Following [2], our similarity score of an incomplete testing video $\mathcal{V}^\delta$ belonging to a specific class $C \in \{1, 2 \cdots, K\}$ is defined as:

$$\begin{aligned} S(C, \mathcal{V}^\delta, l_\mathcal{V}) &= \textstyle\sum_{(f_i, l_i) \in \mathcal{V}^\delta} p(c_i = C, l_\mathcal{V}, f_i, l_i) \\ &\propto \textstyle\sum_{(f_i, l_i) \in \mathcal{V}^\delta} p(c_i = C, l_\mathcal{V} | f_i, l_i), \end{aligned} \tag{2}$$

where $\mathcal{V}^\delta$ refers to the percentage ($\delta \in [0, 100\%]$) of video $\mathcal{V}$ observed, i.e., O in Eq. 1. The prior $p(f_i, l_i)$ in Eq. 2 is assumed to follow a uniform distribution. The probability of $p(c_i = C, l_\mathcal{V} | f_i, l_i)$ can be computed as:

$$\begin{aligned} & p(c_i = C, l_\mathcal{V} | f_i, l_i) \\ = & \sum_{(f_j, s_j, c_j = C) \in \mathcal{D}} p(c_i = C, l_\mathcal{V} | f_j, s_j, c_j = C, f_i, l_i) \\ & \times p(f_j, s_j, c_j = C | f_i, l_i) \\ = & \sum_{(f_j, s_j, c_j = C) \in \mathcal{D}} p(c_i = C, l_\mathcal{V} | s_j, c_j = C, l_i) \\ & \times p(f_j, c_j = C | f_i). \end{aligned} \tag{3}$$

Similar to [2], we made two assumptions for Eq. 3. The first assumption is:

$$p(c_i = C, l_\mathcal{V} | f_j, s_j, c_j = C, f_i, l_i) = p(c_i = C, l_\mathcal{V} | s_j, c_j = C, l_i),$$

referring to the similarity based on the spatial-temporal shifts (white dash line in Fig. 3). We can further compute it as:

$$p(c_i = C, l_\mathcal{V} | s_j, l_i, c_j = C) = \frac{1}{Z} \exp \frac{-((l_i - l_\mathcal{V}) - s_j))^2}{\sigma^2}, \tag{4}$$

where Z is a normalization constant and σ^2 is a bandwidth parameter. The second assumption is that $p(f_j, s_j, c_j =$

$C|f_i, l_i) = p(f_j, c_j = C|f_i)$, which serves as a weight based on feature description for each matched interest point pair $(f_j, s_j, c_j) \in \mathcal{D}$ and $(f_i, l_i) \in \mathcal{V}$.

To reduce the computational cost caused by the enumeration of all the interest point pairs in Eq. 3, we propose a new random forest structure, called *Multi-class Balanced Random Forest* (MBRF), in the next section. With the help of MBRF, we only need to focus on the interest point pairs which fall into the same leaf.

3. MATCHING AND PREDICTING

Random forest has been widely used in many multimedia applications because it has superior performance and fast computational speed. However, for action recognition problems, there are two challenges we need to address. First, since our training data is usually unbalanced, i.e., the number of negative videos will be significantly larger than the number of positive videos. This would easily lead to unbalanced trees, resulting in low discriminative ability and low matching accuracy. Besides, for the human activity prediction problem, one usually needs to recognize multiple categories of actions. It is desirable to develop an algorithm that is scalable to the number of activity classes. Instead of building one-versus-all random forest for each category as in [3], we use a multi-class random forest that saves a lot of computation and storage.

Given the training data, $\mathcal{D} = \{(f_j, s_j, c_j), j = 1, 2, \cdots, N_\mathcal{D}\}$ where f_j is described with Histogram of Gradient (HoG) and Histogram of Flow (HoF), we construct N_T trees as follows. For each node, we choose one of the two splitting measures with equal probability:

- Distribution based measure: to ensure the tree balance and model the underlying data distribution.
- Entropy based measure: to ensure the discriminative ability of the trees.

N_h hypotheses will be generated for each node based on the selected splitting measure. For each hypothesis with the distribution based measure, we randomly select two dimension indexes τ_1 and τ_2 (either from HoG or HoF part). The variance of the training data on the two dimensions can be computed:

$$Var_{(\tau_1,\tau_2)} = \sum_{j=1}^{N_\mathcal{D}} ((f_j(\tau_1) - f_j(\tau_2)) - \mu_{(\tau_1,\tau_2)})^2, \quad (5)$$

where $\mu_{(\tau_1,\tau_2)} = \frac{1}{N_\mathcal{D}} \sum_{j=1}^{N_\mathcal{D}} (f_j(\tau_1) - f_j(\tau_2))$.

Based on the N_h hypotheses, we select the one with the largest variance $Var_{(\tau_1,\tau_2)}$ and the corresponding $\mu_{(\tau_1,\tau_2)}$ is used as the splitting threshold. This distribution based splitting measure has two benefits. On one hand, it can make the trees balanced since the two child nodes after splitting are usually of the similar size. On the other hand, this can be considered as the data distribution modeling. Consider the extreme case when all the nodes are split with distribution based measure, the tree constructing process can be considered as a clustering step, with each leaf as a word in a vocabulary (BoW). The difference between our random forest with the BoW is that we have multiple trees and the variance of the estimation error can be reduced.

Another splitting measure is entropy based measure. Similarly, we generate a set of hypotheses. For each hypothesis, two random numbers for feature dimension indexes are first generated, τ_1 and τ_2. A small jitter value, ξ, is randomly generated as well. We set splitting threshold γ as:

$$\gamma = \frac{1}{N_\mathcal{D}} \sum_{j=1}^{N_\mathcal{D}} (f_j(\tau_1) - f_j(\tau_2)) + \xi. \quad (6)$$

The node will be split into two child nodes based on the threshold γ. For the current node, we can compute the entropy as

$$\begin{aligned} E(\tau_1, \tau_2, \gamma) \quad &= \tfrac{1}{|F_l|} \textstyle\sum_{C=1}^{K} -p_C(F_l) log(p_C(F_l)) \\ &+ \tfrac{1}{|F_r|} \textstyle\sum_{C=1}^{K} -p_C(F_r) log(p_C(F_r)), \end{aligned} \quad (7)$$

where

$$\begin{aligned} F_l &= \{j : f_j(\tau_1) - f_j(\tau_2) < \gamma\} \\ F_r &= \{j : f_j(\tau_1) - f_j(\tau_2) \geq \gamma\}, \end{aligned} \quad (8)$$

and

$$p_C(F_l) = \frac{1}{|F_l|} \sum_{j \in F_l} I(c_j = C), \quad (9)$$

is the probability of samples belonging to the category c in the F_l set. $I(x)$ is an identity function. We can define $p_C(F_r)$ in a similar way. By choosing the hypothesis with the smallest entropy, we can increase the discriminative ability of our trees.

The above process is repeated until the predefined maximum tree depth is reached or the number of feature points in a node is smaller than a pre-defined number. We name this tree structure as Multi-class Balanced Random Forest. Now let us revisit Eq. 3. The weight for the interest point pair can be computed as:

$$p(f_j, c_j = C|f_i) = \frac{1}{|N_T|} \sum_{t=1}^{N_T} \frac{1}{|L_t|} \sum_{j \in L_t} I(c_j = C), \quad (10)$$

where L_t refers to the leaf node in tth tree which both training STIP j and testing STIP i fall in. Rather than enumerating all the interest point pairs between training data and testing video, the computational cost can be significantly reduced by traversing our MBRF and a small subset of interest points from the training data $\mathcal{D}$ will be found with positive weight $p(f_j, c_j = C|f_i)$ for each testing interest point i.

The final score for the segmented testing video is the accumulation of all the votes based on Eq. 2. The category of the video given observation $\mathcal{V}^\delta$ is then determined by:

$$C^* = arg \max_C S(C, \mathcal{V}^\delta, l_\mathcal{V}). \quad (11)$$

With more observations δ provided, more matched pairs will be found and the score will be increased. Thus, the results will be further refined.

In our method, the scale (both spatial and temporal) variations are ignored based on two reasons. First, since our STIP feature already encodes the scale information, the matched STIP pair i and j share the similar activity scale. Second, we add a smooth kernel in Eq. 4 for each vote so that the small scale variations can be well handled.

4. EXPERIMENTS

We choose UT-Interaction dataset [7] to evaluate our algorithm for the following two reasons. First, the dataset is

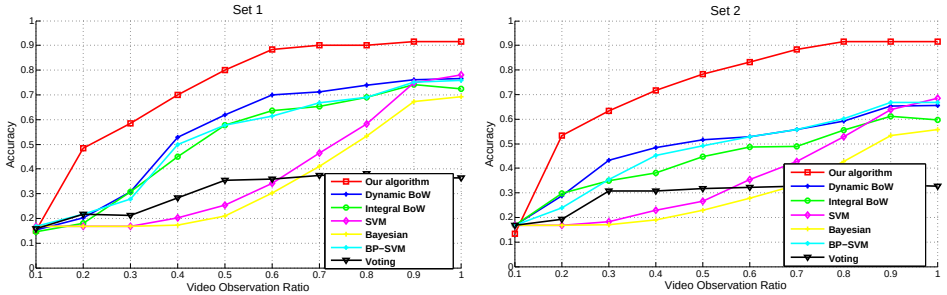

Figure 4: Human activity prediction on UT-Interaction dataset. (Set 1: Left; Set 2: Right)

recorded under the realistic surveillance environment. Second, the activities that a surveillance system is interested in predicting such as shoplifting are usually non-periodic and instantaneous, similar to the six activities in UT-interaction dataset. UT-interaction dataset contains two scenes with 60 videos each. The six types of activities are: *handshaking, pushing, punching, pointing, kicking*, and *hugging*.

The same setting as [6] (Leave-one sequence-out cross validation) is used to evaluate our algorithm. Fig. 4 shows the results compared with the other algorithms on set 1 and set 2, respectively. The following algorithms are compared: *dynamic BoW[6], integral BoW[6], SVM based on Cuboid features, Bayesian classifiers with Gaussian models, BP-SVM: constructing a set of SVMs for each observation level, Voting: a basic voting-based approach that casts a probabilistic vote for each cuboid feature.* The details of these algorithms can be referred to [6].

According to Fig. 4, we obtain on average 20% performance gains over the state-of-the-art techniques. Remarkably, with only 60% observations, our algorithm achieves over 80% accuracy on both set 1 and set 2. This demonstrates that our algorithm is well suited for activity prediction with incomplete observations. Table 1 compares our results with the state-of-the-arts on the UT-Interaction dataset. The results for the first five rows are slightly different from the results in Fig. 4 because Fig. 4 shows the average results with 20 runs of clustering while Table 1 is the best result among 20 runs. We can see that, in both the case of half observations and the case of full observations, our algorithm significantly outperforms the state-of-the-art techniques.

Method	half observation	full observation
Integral BoW [6]	65%	81.7%
Dynamic BoW [6]	70%	85 %
Cuboid + Bayesian [6]	25%	71.7%
Cuboid + SVMs [7]	31.7%	85%
BP-SVM [8]	-	83.3%
Pose 'Doublet' [12]	-	79.17%
Mid-level [4]	-	78.2%
Our proposed	**80%**	**91.7%**

Table 1: Comparison of classification results on UT-Interaction.

In our experiments, the number of trees is set to 100 and the tree depth is set to 15. The feature (STIP) extraction code is downloaded from the author's website [1]. Excluding the feature extraction cost, our algorithm runs in real-time on a normal desktop PC.

5. CONCLUSION

In this paper, we presented a simple yet surprisingly effective solution for human activity prediction problem. Spatial-temporal implicit shape model is proposed to capture the spatio-temporal structure of local features. Matching between the testing and training video is effectively and efficiently solved with our proposed multi-class balanced random forest, which makes a good trade-off between the discriminative ability and tree balance. Besides, our MBRF models all classes simultaneously and therefore is scalable to multi-class prediction. Experimental results show that our algorithm significantly outperforms the state-of-the-arts. In the future work, we plan to handle the activity prediction problem on unsegmented videos.

Acknowledgement

This work was supported in part by the Nanyang Assistant Professorship (SUG M58040015) to Dr. Junsong Yuan.

6. REFERENCES

[1] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, pp. 107–123, 2005.

[2] B. Leibe, A. Leonardis, B. Schiele, "Robust Object Detection with Interleaved Categorization and Segmentation," *IJCV,* 77(1-3), 259-289, 2007.

[3] G. Yu, A. Norberto, J. Yuan, Z. Liu, "Fast Action Detection via Discriminative Random Forest Voting and Top-K Subvolume Search," *IEEE Trans Multimedia,* 2011.

[4] F. Yuan, V. Prinet, J. Yuan, "Middle-Level Representation for Human Activities Recognition: the Role of Spatio-temporal Relationships," *ECCV Workshop on Human Motion,* 2010.

[5] P. Scovanner, S. Ali, M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *ACM Multimedia*, 2007.

[6] M. S. Ryoo, "Human Activity Prediction : Early Recognition of Ongoing Activities from Streaming Videos," in *ICCV*, 2011.

[7] M.S. Ryoo, C. Chen, J. Aggarwal, "An overview of contest on semantic description of human activities (SDHA)," *SDHA,* 2010.

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

[9] G. Yu, J. Yuan, Z. Liu, "Propagative Hough Voting for Human Activity Recognition," *ECCV*, 2012.

[10] G. Zhu, M. Yang, K. Yu, W. Xu, "Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor," *ACM Multimedia*, 2009.

[11] G. Yu, J. Yuan, Z. Liu, "Real-time Human Action Search using Random Forest based Hough Voting," *ACM Multimedia*, 2011.

[12] S. Mukherjee, S.K. Biswas, D.P. Mukherjee, "Recognizing Interaction Between Human Performers Using 'Key Pose Doublet'," *ACM Multimedia*, 2011.

[13] G. Yu, J. Yuan, Z. Liu, "Unsupervised Random Forest Indexing for Fast Action Search," *CVPR*, 2011.