

Speeding up Spatio-Temporal Sliding-Window Search for Efficient Event Detection in Crowded Videos

Junsong Yuan
EECS department
Northwestern University
2145 Sheridan Road
Evanston, IL, USA
j-yuan@u.northwestern.edu

Zicheng Liu
Microsoft Research
One Microsoft Way
Redmond, WA, USA 98052
zliu@microsoft.com

Ying Wu
EECS department
Northwestern University
2145 Sheridan Road
Evanston, IL, USA
yingwu@ece.northwestern.edu

Zhengyou Zhang
Microsoft Research
One Microsoft Way
Redmond, WA, USA 98052
zhang@microsoft.com

ABSTRACT

Despite previous successes of sliding window-based object detection in images such as [6], searching desired events in the volumetric video space is still a challenging problem, partially because the pattern search in spatio-temporal video space is much more complicated than that in spatial image space. Without knowing the location, temporal duration, and the spatial scale of the event, the search space for video events is prohibitively large for exhaustive search. To reduce the search complexity, we propose a heuristic branch-and-bound solution for event detection in videos. Unlike existing branch-and-bound method which searches for an optimal subvolume before comparing its detection score against the threshold, we aim at directly finding subvolumes whose scores are higher than the threshold. In doing so, many unnecessary branches are terminated much earlier, thus the search speed can be much faster. To validate this approach, we select three human action classes from the KTH dataset for training while testing with our own action dataset which has clutter and moving backgrounds as well as large variations in lighting, scale, and performing speed of actions. The experiment results show that our technique dramatically reduces computational cost without significantly degrading the quality of the detection results.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Search Process*; H.2.8 [Database Management]: Database Applications—*Image Database, Data Mining*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

EiMM'09, October 23, 2009, Beijing, China.

Copyright 2009 ACM 978-1-60558-754-7/09/10 ...\$10.00.

General Terms

Algorithms, Performance

Keywords

event detection, spatio-temporal pattern, video pattern search, sliding window

1. INTRODUCTION

Event detection in video has been an interesting yet challenging problem. It has a wide range of applications including video surveillance, medical diagnosis and training, robotics, smart rooms, video indexing, and human computer interaction. Unlike event categorization [19] [22] [21], which classifies a whole video clip or a collection of images into one of the predefined event classes, event detection is a more challenging task because it requires an accurate spatial and temporal localization of the event in the video.

Similar to the use of sliding windows for object detection in images, event detection is to search a spatio-temporal window, *i.e.* a subvolume that contains the target event in video [5]. Despite previous successes of sliding window-based object detection [6], locating desired events in video volumetric data is still a challenging problem. First of all, the search complexity in the spatio-temporal video space is much higher than that of searching for objects in the image space. Besides the spatial location and scale of the event, we also need to determine its temporal location and duration. Such a search space in volumetric video data is prohibitively large for exhaustive search. Moreover, events such as actions often exhibit tremendous amount of intra-pattern variations caused by changes in performing speed, scale, lighting, backgrounds, and possibly partial occlusions. It is extremely difficult to handle all of these variations in detection.

Recently, an event detection technique that addresses the above problems is proposed in [20], where it is applied to multi-class human actions detection in videos. In their method, a video sequence is represented by a collection of spatio-temporal interest points (STIPs) [14] [7]. Based on the naive-Bayes assumption [1], each STIP casts a positive or negative-valued vote independently for a specific action class,

where the voting score is based on its mutual information with respect to the action class. Event detection is formulated as the problem of searching for the spatio-temporal subvolume that has the maximum total votes. In such a case, the mutual information between the detected subvolume and the action class is maximized. To handle the extremely large search space in video, they proposed a branch and bound algorithm which decouples the temporal and spatial spaces and applies different search strategies on them for efficient search.

In this work, we present a new technique to further improve the efficiency of the subvolume search in videos. For detection applications, previous branch-and-bound approaches such as [6] [20] search for a unique subvolume of maximum score, then compare its detection score against the threshold to determine whether it is a valid detection or not. In comparison, we aim at directly finding subvolumes whose scores are higher than the threshold. Under the guidance of the detection threshold, the search process can be much faster by terminating many unnecessary branches earlier during the search process of branch-and-bound. It brings a much faster search, without significantly degrading the quality of the detection results. To validate the proposed method, we select 3 types of human actions and perform a cross-dataset experiment for multiclass action detection, where the training and test data are from different sources. The test video exhibits large lighting variations, clutter and moving backgrounds, and partial occlusions of the actions. Our experimental results demonstrate the effectiveness and efficiency of our method.

2. RELATED WORK

There has been increasing interests in characterizing human actions using a set of spatio-temporal interest points (STIPs) [14] [7]. Feature vectors are extracted from local volumes centered at a number of STIPs and action classifiers are then built upon the feature vectors. This approach is partly motivated from the progress in image classification and object recognition where bag-of-feature approaches have become very popular [1]. For action recognition, such a bag-of-feature approach has the advantage that it does not require tracking or foreground-background separation.

Another type of previous approaches treat actions as spatio-temporal templates, then the action can be detected through finding the matches of spatio-temporal template [15] [5] [12] [2]. To improve the detection performance, both shape and motion information are applied for action detection [11] [17]. In [18], a generative model is learned by using both semantic and structure information for action recognition and detection. In [3], a successive convex matching scheme is proposed for action detection. Actions are detected in video by matching the time-coupled posture sequences to video frames.

Some other methods apply discriminative approach for action detection. In [4], Haar features are extended to 3-dimensional space and boosting is applied by integrating these features for final classification. In [5], a given video is segmented into a few subvolumes. To detect actions, an action template is matched by searching among the over-segmented regions.

Our method belongs to the category of bag-of-feature approaches. Even though quite some progress has been made on action classification [10] [16] [9] [10], much less work has

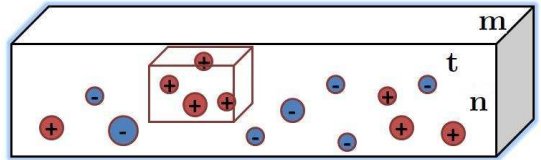


Figure 1: Event detection as spatio-temporal subvolume search. The highlighted subvolume has the maximum mutual information toward the specific event class. Each circle represents a spatio-temporal feature point which contributes a vote based on its own mutual information. The whole video volume is of size $m \times n \times t$, where $m \times n$ is the image size and t is the temporal length.

been reported on action detection with STIPs. In our framework, actions are treated as spatio-temporal objects which are characterized as 3-dimensional volumetric data. Each action can be represented as a collection of spatio-temporal invariant features.

3. EVENT DETECTION MODEL

Given a video sequence $\mathcal{V} \in \mathbb{R}^{m \times n \times t}$, our goal is to identify where (spatial location in the image) and when (temporal location) a specific event occurs, namely to find a spatio-temporal subvolume (3-dimensional subvolume) $V^* \subseteq \mathcal{V}$, such that it has the maximum detection score:

$$V^* = \arg \max_{V \subseteq \mathcal{V}} f(V) = \arg \max_{V \in \Lambda} f(V),$$

where $f(V)$ is the detection score for the spatio-temporal subvolume V , and Λ denotes the candidate set of all valid subvolumes in \mathcal{V} .

According to [20], we illustrate the subvolume search for event detection in Fig. 1. Suppose the target video \mathcal{V} is of size $m \times n \times t$. The optimal solution is denoted by:

$$V^* = t^* \times b^* \times l^* \times r^* \times s^* \times e^*,$$

which has 6 parameters to be determined, where $t^*, b^* \in [0, m]$ denote the top and bottom positions, $l^*, r^* \in [0, n]$ denote the left and right positions, and $s^*, e^* \in [0, t]$ denote the start and end positions. As a counterpart of the bounding-box based object detection, the solution V^* is the bounding volume that has the highest detection score of the target action.

3.1 Evaluation of Detection Score

It is of essential importance to evaluate the detection score $f(V)$ appropriately. First of all, we characterize an action as a collection of spatio-temporal interest points (STIPs) [7], and apply two types of features to describe each STIP [8]: histogram of gradient (HOG) and histogram of flow (HOF). HOG is the appearance feature and HOF is the motion feature. As STIPs are locally invariant features, they are relatively robust to action variations due to the changes in performing speed, scale, lighting condition and cloth. We denote a video sequence by $\mathcal{V} = \{d_i\}$, where $d \in \mathbb{R}^N$ is a feature vector describing a STIP.

We consider a multi-class problem and let $\mathbf{C} = \{1, 2, \dots, C\}$ be the class label set. We follow the naive-Bayes mutual information maximization (NBMIM) proposed in [20] to evaluate the detection score $f(V)$. Assuming the independence among the STIP and an equal prior, *i.e.* $P(\mathbf{C} = c) = \frac{1}{C}$,

the detection score for each point $d \in \mathcal{V}$ can be evaluated as following according to [20]:

$$s^c(d) = \log \frac{C}{1 + \frac{P(d|C \neq c)}{P(d|C=c)}(C-1)}, \quad (1)$$

where $\frac{P(d|C \neq c)}{P(d|C=c)}$ is the likelihood ratio test and can be calculated through non-parametric density estimation. For the C -class action detection, each point d has C detection scores toward C classes. For a specific class c , the detection score of a subvolume V is the summation of all point scores within the subvolume:

$$f(V) = \sum_{d \in V} s^c(d),$$

where $s^c(d) > 0$ votes for class c , while $s^c(d) < 0$ votes for negative class.

4. OUR DETECTION ALGORITHM

The speed of the event detection, or in general video pattern search, is extremely critical to handle large video dataset. As the search space of V^* contains 6 parameters, it is extremely time consuming to apply an exhaustive search. To speed up the search process, we follow the branch-and-bound solution proposed in [6] [20], but aims at a further improvement on the efficiency. In the following, we first review the upperbound estimation of $f(V^*)$ for efficient pruning in Section 4.1, followed by a summary of conventional branch and bound search in Section 4.2. Then we propose the new detection algorithm in Section 4.3.

4.1 Upperbound Estimation of the Detection Score

We consider a 3-dimensional subvolume parameter space denoted by $\mathbb{V} = [\mathbb{W}, \mathbb{T}]$, where

$$\mathbb{W} = [0, n] \times [0, n] \times [0, m] \times [0, m]$$

is the 4-dimensional spatial parameter space and

$$\mathbb{T} = [0, t] \times [0, t]$$

is the 2-dimensional temporal parameter space. Let the optimal solution within \mathbb{V} be

$$V^* = \arg \max_{V \subseteq \mathbb{V}} f(V)$$

and suppose $D_t > 0$ is the detection score. The parameter space \mathbb{V} is a *valid* solution space if and only if $f(V^*) \geq D_t$.

Our purpose is to find an upperbound estimation of V^* for a parameter space \mathbb{V} , such that $\hat{f}(\mathbb{V}) \geq f(V^*)$. With the help of $\hat{f}(\mathbb{V})$, we can safely discard \mathbb{V} when $\hat{f}(\mathbb{V}) < D_t$, because \mathbb{V} is not a valid space. To enable efficient pruning, we require the upperbound estimation $\hat{f}(\mathbb{V})$ is tight and efficient.

Fixing a spatial window W , we define its optimal score along the temporal line as:

$$F(W) = \max_{T \subseteq \mathbb{T}} f(W, T), \quad (2)$$

where $f(W, T) = \sum_{d \in W \times T} s(d)$ denotes the detection score of the subvolume $V = [W, T]$. Then for each pixel $i \in W$, we define the following two quantities:

$$F(i) = \max_{T \subseteq \mathbb{T}} f(i, T), \quad (3)$$

and

$$G(i) = \min_{T \subseteq \mathbb{T}} f(i, T). \quad (4)$$

Let $F^+(i) = \max(F(i), 0)$ and $G^-(i) = \min(G(i), 0)$, given a spatial parameter space $\mathbb{W} = \{W : W_{min} \subseteq W \subseteq W_{max}\}$, we have the upperbound estimation of $\hat{f}(V^*)$ [20].

$$\hat{f}(\mathbb{V}) = \min\{\hat{F}_1(\mathbb{W}), \hat{F}_2(\mathbb{W})\} \geq f(V^*),$$

where

$$\hat{F}_1(\mathbb{W}) = F(W_{min}) + \sum_{i \in W_{max}, i \notin W_{min}} F^+(i),$$

and

$$\hat{F}_2(\mathbb{W}) = F(W_{max}) - \sum_{i \in W_{max}, i \notin W_{min}} G^-(i).$$

We can apply $\hat{f}(\mathbb{V})$ as the upperbound estimation in the branch and bound solution. The estimation of $\hat{f}(\mathbb{V})$ is of only linear complexity toward the video length, thus it is efficient compared with finding the $f(V^*)$ in a parameter space \mathbb{V} .

4.2 Branch-and-Bound Solution

A new search method called Spatio-temporal branch-and-bound search (STBB) is presented in [20] to search the video space. Instead of directly applying branch-and-bound in the 6D parameter space, it decomposes the parameter space into two subspaces: (1) 4D spatial parameter space and (2) 2D temporal parameter space. Different search strategies are performed in the two subspaces \mathbb{W} and \mathbb{T} . It optimizes the parameter alternately between \mathbb{W} and \mathbb{T} .

Alg.0: spatio-temporal brand-and-bound search [20]

Require: video $\mathcal{V} \in \mathbb{R}^{m \times n \times t}$

Require: quality bounding function \hat{F} (see text)

Ensure: $V^* = \arg \max_{V \subseteq \mathcal{V}} f(V)$

set $\mathbb{W} = [T, B, L, R] = [0, n] \times [0, n] \times [0, m] \times [0, m]$

get $\hat{F}(\mathbb{W}) = \min\{\hat{F}_1(\mathbb{W}), \hat{F}_2(\mathbb{W})\}$

push $(\mathbb{W}, \hat{F}(\mathbb{W}))$ into empty priority queue P

set current best solution $\{W^*, F^*\} = \{W_{max}, F(W_{max})\}$;

repeat

retrieve top state \mathbb{W} from P based on $\hat{F}(\mathbb{W})$

if $(\hat{F}(\mathbb{W}) > F^*)$

split $\mathbb{W} \rightarrow \mathbb{W}^1 \cup \mathbb{W}^2$

CheckToUpdate($\mathbb{W}_1, W^*, F^*, P$);

CheckToUpdate($\mathbb{W}_2, W^*, F^*, P$);

else

$T^* = \arg \max_{T \subseteq [0, t]} f(W^*, T)$;

return $V^* = [W^*, T^*]$.

function **CheckToUpdate**(\mathbb{W}, W^*, F^*, P)

Get W_{min} and W_{max} of \mathbb{W}

if $(F(W_{min}) > F^*)$

update $\{W^*, F^*\} = \{W_{min}, F(W_{min})\}$;

if $(F(W_{max}) > F^*)$

update $\{W^*, F^*\} = \{W_{max}, F(W_{max})\}$;

if $(W_{max} \neq W_{min})$

get $\hat{F}(\mathbb{W}) = \min\{\hat{F}_1(\mathbb{W}), \hat{F}_2(\mathbb{W})\}$

if $\hat{F}(\mathbb{W}) > F^*$

push $(\mathbb{W}, \hat{F}(\mathbb{W}))$ into P

4.3 Improved Branch-and-Bound Solution for Detection

As discussed before, conventional branch-and-bound solution can be further sped up. Instead of searching for an optimal V^* with maximum detection score, we can safely terminate the search if none of the candidates satisfies $\hat{f}(\mathbb{V}) \geq D_t$. In such a case, it is guaranteed that $f(V^*) < D_t$, thus \mathbb{V} is not a valid parameter space. Furthermore, if a subvolume with valid detection score $f(V) \geq D_t$ is found during the search, we can quickly finalize the detection based on the current solution, instead of keeping looking for V^* . Incorporating the above two heuristics, we present the improved branch and bound algorithm in Alg.1.

Compared with previous methods in [6] [20], during each search iteration, we retrieve an upper bounded estimation $\hat{F}(\mathbb{W})$ from the heap. If $\hat{F}(\mathbb{W}) < D_t$, we directly reject the whole video sequence \mathcal{V} , since no V^* can meet the detection threshold. This strategy largely speeds up the scanning of negative video clips which do not contain the target action. Moreover, at each search iteration, we also keep track of the current best score F^* . When $F^* \geq D_t$, it is guaranteed that there exists a valid detection in the corresponding parameter space \mathbb{W} that contains F^* . In such a case, we speed up by limiting the rest of the search within \mathbb{W} only. In other words, instead of searching for the optimal $f(V^*)$, we are satisfied with a qualified detection $f(\tilde{V}) > D_t$, although it is possible that $f(\tilde{V}) < f(V^*)$.

Alg.1: Accelerated Branch and Bound for Detection

Input: video $\mathcal{V} \in \mathbb{R}^{m \times n \times t}$, detection threshold D_t

Output: subvolume $\tilde{V} \subseteq \mathcal{V}$, s.t. $f(\tilde{V}) \geq D_t$ (or $\tilde{V} = \emptyset$)

set $\mathbb{W} = [T, B, L, R] = [0, n] \times [0, n] \times [0, m] \times [0, m]$

get $\hat{F}(\mathbb{W}) = \min\{\hat{F}_1(\mathbb{W}), \hat{F}_2(\mathbb{W})\}$

push $(\mathbb{W}, \hat{F}(\mathbb{W}))$ into empty priority queue P

set current best solution $\{W^*, F^*\} = \{W_{max}, F(W_{max})\}$;

repeat

retrieve top state \mathbb{W} from P based on $\hat{F}(\mathbb{W})$

if $\hat{F}(\mathbb{W}) < D_t$

return $\tilde{V} = \emptyset$

if $(\hat{F}(\mathbb{W}) > F^*)$

split $\mathbb{W} \rightarrow \mathbb{W}^1 \cup \mathbb{W}^2$

CheckToUpdate($\mathbb{W}_1, W^*, F^*, P$);

CheckToUpdate($\mathbb{W}_2, W^*, F^*, P$);

else

$T^* = \arg \max_{T \in [0, t]} f(W^*, T)$;

return $\tilde{V} = [W^*, T^*]$.

function CheckToUpdate(\mathbb{W}, W^*, F^*, P)

if $(F^* \geq D_t)$

clear priority queue P

push $(\mathbb{W}, \hat{F}(\mathbb{W}))$ into empty priority queue P

else

Get W_{min} and W_{max} of \mathbb{W}

if $(F(W_{min}) > F^*)$

update $\{W^*, F^*\} = \{W_{min}, F(W_{min})\}$;

if $(F(W_{max}) > F^*)$

update $\{W^*, F^*\} = \{W_{max}, F(W_{max})\}$;

if $(W_{max} \neq W_{min})$

get $\hat{F}(\mathbb{W}) = \min\{\hat{F}_1(\mathbb{W}), \hat{F}_2(\mathbb{W})\}$

if $\hat{F}(\mathbb{W}) > F^*$

push $(\mathbb{W}, \hat{F}(\mathbb{W}))$ into P

5. EXPERIMENTS

5.1 Efficiency Test

To validate the efficiency of our proposed algorithm, we select the first five video sequences of two-hand waving in the CMU action dataset [5]. Each sequence contains a two-hand wave action and we want to detect a subvolume that covers the action. The detection threshold is selected as $D_t = 10$, which does not affect the efficiency of our approach. The detected subvolume will be compared with the ground truth subvolume.

In Table 1, we compare our proposed Alg.1 with the conventional branch-and-bound method in Alg.0, which finds the unique optimal subvolume of maximum score. W^* is the spatial window containing left, right, top, and bottom parameters. T^* includes the start and end frames. It shows that detection results of Alg.1 are close to those of conventional branch-and-bound, in terms of detection scores and subvolume parameters. However, the search speed can be tens of times faster than the conventional approach in terms of the number of branches. We implement Alg.1 using MATLAB and it can process videos of resolution 160×120 at around 5-30 frames per second, without including the extraction of STIPs and calculation of voting score $s(d)$.

The overall cost of our method is from three aspects: (1) extraction of STIPs; (2) calculation of voting scores and (3) search of the spatio-temporal subvolume. First, for our videos of resolution 160×120 , the detection speed of the STIPs is 4-8 frames per second using the binary code provided by [8]. Second, by using LSH for efficient nearest neighbor search, the query time of each STIP is only 10 to 20 milliseconds. As each frame contains 5-10 STIPs on average, the processing time is 5-20 frames per second. Finally, for the efficiency purpose, we apply the approximated strategy mentioned above in searching for the spatio-temporal bounding box in our experiment. This leads to an efficient MATLAB implementation at around 5-30 frames per second. Overall, our method has the potential to be implemented in real-time.

5.2 Multi-class action detection

To test our method on multi-class action detection, we select 3 classes of actions: boxing, hand waving and hand clapping from the standard KTH action dataset [13]. We use KTH dataset as the training data while using our own dataset for testing. The KTH dataset contains 25 sequences performed by 25 subjects for each action type. The data are captured with clean background, but it contains actions with different scales, view point changes, and style variations. It contains both indoor and outdoor scenes. To better distinguish our target actions from the clutter and moving background, we introduce the walking class in KTH dataset (25 sequences) as the background class. Together with the background class, we perform a 4-class classification in calculating the score in Eq. 1. However, our detection task does not include walking. For each of the three action classes, its negative STIP set includes those from the background class (walking), as well as the other two action classes.

The test dataset are captured and labeled by ourselves. It contains 10 video sequences and has in total 44 actions: 10 hand clapping, 16 hand waving, and 18 boxing, performed by 8 subjects. Each sequence contains multiple types of actions. There are both indoor and outdoor scenes. All of the

	W^*	T^*	score	# of branches
V1: Alg.0	55 97 23 54	442 553	16.21	206933
V1: Alg.1	55 97 23 52	442 546	15.97	4549
V2: Alg.0	61 122 20 38	673 858	37.39	67281
V2: Alg.1	60 122 20 39	673 858	37.36	4668
V3: Alg.0	72 118 22 71	11 700	89.01	71594
V3: Alg.1	82 114 23 73	10 705	85.21	2275
V4: Alg.0	73 112 23 77	420 1083	73.42	63076
V4: Alg.1	77 108 23 78	420 1083	70.93	2363
V5: Alg.0	18 144 7 114	418 451	46.50	315770
V5: Alg.1	41 151 7 114	419 451	45.36	133027

Table 1: The comparison of search results between our Alg.1 and conventional branch-and-bound in Alg.0. We test our method on the first five sequences in the CMU hand-waving dataset. It shows our proposed method in Alg.1 can be up to 20 times faster than that of Alg.0, while still achieve similar performance as the optimal solution. We manually select the detection threshold $D=10$. However, the efficiency of our approach does rely on this parameter.

video sequences are captured with clutter and moving backgrounds, where human tracking and detection is very difficult. Each video is of low resolution 160×120 and frame rate 15 frames per second. Their lengths are between 34 to 76 seconds. To evaluate the performance, we manually label a spatio-temporal bounding box for each action. We apply similar measurement proposed in [5] but with a relative loose criterion. A detected action is regarded as correct if at least 1/4 of the volume size overlaps with the ground truth label. On the other hand, an action is regarded as retrieved if at least 1/8 of its volume size overlaps with a detection. We apply the precision and recall scores to measure the performance.¹ To filter noisy detections, we require a valid detection lasts more than 20 frames while less than 200 frames.

Because only a very small portion of pixels $d \in \mathcal{V}$ are associated with STIPs, we put a constant negative prior $s(d_\theta) < 0$ to the rest of pixels which do not associate with any STIP. In other word, if a pixel does not associate with a STIP, we tend to classify it into the background class and it will vote a negative score toward any positive action class. With such a negative prior, we put a penalty on detections of large spatio-temporal scales, *i.e.* large volume size. Another advantage is that this brings a unique maximum solution when searching the subvolume with maximum score.

In Fig. 5.2, we show the precision and recall curves for three class actions, by changing the detection threshold D_t from 1 to 20. We also compare two difference choices of $s(d_\theta) = -5 \times 10^{-5}$ and $s(d_\theta) = -1 \times 10^{-4}$. It shows the negative penalty $s(d_\theta) = -1 \times 10^{-4}$ performs slightly better. Among the three types of actions, hand waving gives much better detection performance than clapping and boxing. Boxing is the most challenging action class.

6. CONCLUSION AND FUTURE WORK

We have presented a data-driven approach for event detection in crowded videos. A video event is characterized as a spatio-temporal point collection and we apply the naive-Bayes based mutual information to measure the detection score. To efficiently search the spatio-temporal video space

¹precision = # correct detect / # total detect
recall = # correct detect / # total action.

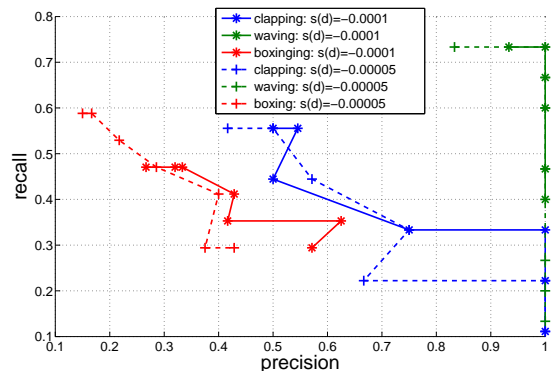


Figure 2: Detection performance of three types of action: hand clapping, hand waving and boxing.

for event detection, an accelerated branch-and-bound solution is proposed by using a heuristic search strategy. The proposed method dramatically reduces search complexity without significantly degrading the quality of the detection results. To test our method, we treat human actions as specific events. Our proposed detection method does not rely on human tracking and detection, and can handle scale variations, clutter and moving background, and even partial occlusions. To evaluate the action detection performance, a new action dataset with ground truth labeling is provided and tested.

Although we only use human actions as concrete examples of events, the proposed video pattern search approach provides a general and efficient solution to other types of event detection. Our future work includes a further improvement of the detection accuracy and the extension of our method to other video analysis tasks.

7. REFERENCES

- [1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.
- [2] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. IEEE International Conf. on Computer Vision*, 2003.



Figure 3: Examples of detection results. Each row shows an action class: hand waving (first row), hand clapping (second row) and boxing (third row). Each image is a sample frame from the action video. The first column shows the training videos from the KTH dataset. The second to the fifth columns are some detection results. The highlighted bounding boxes correspond to the detected spatial window W^* and we apply different colors to distinguish different action types: clapping (turquoise), waving (magenta), and boxing (yellow). The final column gives the miss detection examples, where the bounding boxes are ground truth labeling: clapping (red), waving (green), and boxing (blue).

- [3] H. Jiang, M. S. Drew, and Z.-N. Li. Successive convex matching for action detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.
- [4] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proc. IEEE International Conf. on Computer Vision*, 2005.
- [5] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *Proc. IEEE International Conf. on Computer Vision*, 2007.
- [6] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.
- [7] I. Laptev. On space-time interest points. *Intl. Journal of Computer Vision*, 2005.
- [8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.
- [9] Z. Li, Y. Fu, S. Yan, and T. S. Huang. Real-time human action recognition by luminance field trajectory analysis. In *Proc. ACM Multimedia*, 2008.
- [10] J. Liu and M. Shah. Learning human actions via information maximization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.
- [11] P. Natarajan and R. Nevatia. View and scale invariant action recognition using multiview shape-flow models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.
- [12] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.
- [13] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proc. IEEE Conf. on Pattern Recognition*, 2004.
- [14] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc. ACM Multimedia*, 2007.
- [15] E. Shechtman and M. Irani. Space-time behavior based correlation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.
- [16] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.
- [17] S. N. Vitaladevuni, V. Kellokumpu, and L. S. Davis. Action recognition using ballistic dynamics. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.
- [18] S.-F. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.
- [19] D. Xu and S.-F. Chang. Video event recognition using kernel methods with multi-level temporal alignment. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(11):1985–1997, 2008.
- [20] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.
- [21] J. Yuan, J. Luo, H. Kautz, and Y. Wu. Mining gps traces and visual words for event classification. In *Proc. ACM International Conference on Multimedia Information Retrieval*, 2008.
- [22] J. Yuan, J. Luo, and Y. Wu. Mining compositional features for boosting. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.