# Real-time Action Recognition by Spatiotemporal Semantic and Structural Forests

Tsz-Ho Yu
thy23@eng.cam.ac.uk
Tae-Kyun Kim
http://mi.eng.cam.ac.uk/~tkk22
Roberto Cipolla
cipolla@eng.cam.ac.uk

Machine Intelligence Laboratory
Department of Engineering
University of Cambridge
Trumpington Street, Cambridge
CB2 1PZ, UK

**Abstract**

Whereas most existing action recognition methods require computationally demanding feature extraction and/or classification, this paper presents a novel real-time solution that utilises local appearance and structural information. Semantic texton forests (STFs) are applied to local space-time volumes as a powerful discriminative codebook. Since STFs act directly on video pixels without using expensive descriptors, visual codeword generation by STFs is extremely fast. To capture the structural information of actions, so called pyramidal spatiotemporal relationship match (PSRM) is introduced. Leveraging the hierarchical structure of STFs, the pyramid match kernel is applied to obtain robust structural matching, avoiding quantisation effects. We propose the kernel k-means forest classifier using PSRM to perform classification. In the experiments using KTH and the latest UT-interaction data sets, we demonstrate real-time performance as well as state-of-the-art accuracy by the proposed method.

## 1 Introduction

Recognising human actions from videos has been widely studied for applications such as human-computer interaction, digital entertainment, visual surveillance and automatic video indexing. Despite the popularity of the topic in computer vision research, some issues still remain for realising its potentials:

- While *time efficiency* is of vital importance in real-world action recognition systems, current methods seldom take computational complexity into full consideration. State-of-the-art algorithms (*e.g.* [5, 8, 9, 27]) have reported satisfactory accuracies on standard human action data sets. They, however, often resort to computationally heavy algorithms to obtain the good accuracies.

- Action classification with *a short response time* is useful for continuous recognition in human-computer interaction. Typically, a class label is assigned after an entire query video is analysed, or a large lookahead is required to collect sufficient features. In fact,


BMVC 2010 doi:10.5244/C.24.52

as suggested by [22], actions can be recognised from very short sequences called the "snippets".

- *Structural information* is a useful cue for action recognition. The "bag of words"(BOW) has proven a effective model for action recognition owing to its rich description power of local appearance information and its inherent benefits to cope with scale changes, translation and cluttered backgrounds. However, the standard BOW model ignores the spatiotemporal relationships among local descriptors.

Addressing the aforementioned challenges, we present a novel method for human action recognition. The goal of this work is to design a very fast but competitively accurate action recogniser over state-of-the-arts. The major contributions include the followings:

**Efficient Spatiotemporal Codebook Learning:** We extend the use of semantic texton forests [25] (STFs) from 2D image segmentation to spatiotemporal analysis. STFs are ensembles of random decision trees that translate interest points into visual codewords. In our method, STFs perform directly on video pixels without computing expensive local descriptors. As well as being much faster than a traditional flat codebook such as k-means clustering, STFs achieve high accuracy comparable to that of existing approaches.

**Combined Structural and Appearance Information:** We propose a richer description of features, hence actions can be classified in very short video sequences. Building on the work of Ryoo and Aggarwal [19], we introduce pyramidal spatiotemporal relationship match (PSRM). Histogram intersection used in [19] is prune to quantisation errors when the histograms have a large number of bins. Taking the inherent benefit of the hierarchical structure of semantic texton forests, the pyramidal match kernel [9] is employed to alleviate this problem.

**Improved Recognition Performance:** Several techniques are employed to enhance the recognition speed and accuracy. A novel spatiotemporal interest point detector, called V-FAST, is designed based on the FAST 2D corners [18]. A fast and effective classifier, namely k-means forest classifier, is also proposed. The recognition accuracy is improved by adaptively combining PSRM and the bag of semantic texton (BOST) method [25].

The rest of the paper is structured as follows: In section 2, related works are reviewed. In section 3–7, the proposed methods are detailed. Evaluation results are reported and discussed in Section 8 and the conclusion is drawn in Section 9.

## 2 Related Work

State-of-the-art action recognition methods have shown the effectiveness of local appearance-based features: the "bag of words" is a widely used technique in the literature [2, 14, 17, 23, 28]. A codebook is learned to quantise input features into visual codewords. Classification is then performed on the histograms of codewords. Generally, a large-sized codebook is required to obtain high recognition accuracy, yet an oversized codebook leads to high quantisation errors and overfitting. K-means clustering is a popular algorithm for codebook learning. Feature quantization by a large flat codebook such as k-means is, however, computationally heavy. Tree-based codebooks have been explored as an alternative to speed up the feature quantisation. Since Moosmann *et al*. [13], random forests have been increasingly used in many tasks *e.g*. image classification and segmentation [25], owing to good generalisation and efficiency. Similarly, Oshin *et al*. [15] recognise actions by analysing the distribution of

interest points by random ferns. Lin *et al*. [8] used a prototype tree to encode holistic motion-shapes descriptors. Mikolajczyk and Uemura [12] built clustering trees from the centroids obtained by k-means clustering. Hierarchical codebooks enable fast vector quantisations, but the expensive features and classifiers used in [8, 12] make the overall processes still heavy.

Standard bag of word models contain only local appearance information. While structural context could be useful for describing action classes, it is often overlooked in current action recognition methods. Several recent studies have attempted to augment structural information into local appearance features. Scovanner *et al*. [24] employ a two-dimensional histogram to describe feature co-occurrences. Savarese *et al*. [21] propose "correlograms" to measure the similarity of actions globally. Wong *et al*. [28] present the pLSA-ISM model, which is an extension of the probabilistic latent semantic analysis (pLSA) by spatial information. Tran and Sorokin [26] and Zhang *et al*. [30] capture structural information directly by a global shape descriptor. Since these methods [26, 28, 30] encode holistic structures with respect to a reference position *e.g*. the center of ROI (region of interests), they require manual segmentation or computationally-demanding detection of ROI. Structural relationships among individual features are not fully utilised in these methods. Most recently, Ryoo and Aggarwal [19] propose the spatiotemporal relationship match (SRM) which represents structures by a set of pairwise spatiotemporal association rules. Kovashka and Grauman [6] exploit structural information by learning an optimal neighbourhood measure on interest points. Despite of the high accuracies reported, speed and quantisation errors are the major issues due to the flat k-means codebook involved.

The pyramid match kernel (PMK) [4] is widely used in recent image-based object detection and matching studies. PMK exploits multi-resolution histograms. Similar points that do not match at fine resolutions have a chance to match at lower resolutions. Hence, PMK reduces quantisation errors and enhances robustness. Liu and Shah [9] matched interest points in multiple resolutions using PMK and reported improved results, however the features are only matched spatially but not semantically.

Design of interest point detector/descriptor and classifiers also plays an essential role. Just to name a few, the detectors designed by Laptev and Lindeberg [7] and Dollar *et al*. [2] are commonly adopted in existing methods. Both of them are the extensions of two-dimensional Harris corners. To describe interest points, histograms of gradients (HOG) and optical flow are popular in earlier approaches [2, 14, 23]. Scovanner *et al*. [24] proposed a three-dimensional version of Lowe's popular SIFT descriptors [10]. Willems *et al*. [27] used an extended SURF descriptor for action recognition. Some common classifiers used in action recognition include K-NN classifiers, support vector machines and boosting, which are complex to attain sufficient real-time performance.

With increasing interests in practical applications, real-time action recognition algorithms have attained new attentions. For instance, Yeffet and Wolf [29] utilise dense local trinary patterns with a linear SVM classifier. Gilbert *et al*. [3] propose a fast multi-action recognition algorithm by finding reoccurring patterns on dense 2D Harris corners by a data-mining algorithm. Patron-Perez and Reid [16] designed a probabilistic classifier that recognises actions continuously by a sliding window. Bregonzio *et al*. [11] consider actions as clouds of points, and efficient classification is done by analysing histograms of point clusters. The requirement of prior segmentation or long sequences for classification renders the respective methods not responsive.

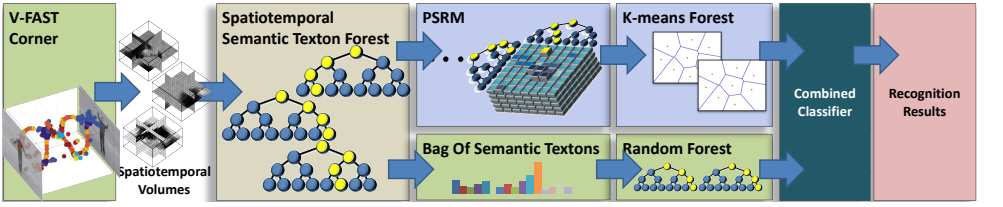


Figure 1: Overview of the proposed approach

# 3 Overview

An overview of the proposed approach is illustrated in figure 1. Firstly, spatiotemporal interest points are localised by the proposed V-FAST detector. Semantic texton forests (STFs) are learned to convert local spatiotemporal patches to visual codewords. Secondly, structural information of human actions is captured by the pyramidal spatiotemporal relationship match (PSRM). Classification is then performed efficiently using a hierarchical k-means algorithm with the pyramid match kernel. The proposed method is adaptively combined with the prior-art that uses the bag of semantic textons (BOST) and random forests as a classifier to further improve the recognition accuracy.

# 4 V-FAST Interest Point Detector

V-FAST (Video FAST) interest points are obtained by extending the FAST corners [13] into a spatiotemporal domain. It considers pixels in three orthogonal Bresenham circles with a radius $r$ on $XY$, $YT$ and $XT$ planes. Similar to FAST, saliency is detected on a plane if there exist $n$ contiguous pixels on the circle which are all brighter than a reference pixel $p(x,y,t)$ plus a threshold $t$, or all darker than $p(x,y,t)-t$. An interest point is detected when the reference pixel shows both spatial ($XY$-plane) and temporal ($XT$-plane or $YT$-plane) saliency. The V-FAST detector gives a dense set of interest points, which enables accurate classification from relatively short sequences. Figure 2 illustrates how interest points are detected using the 42-pixel V-FAST interest point detector with $r = 3$.

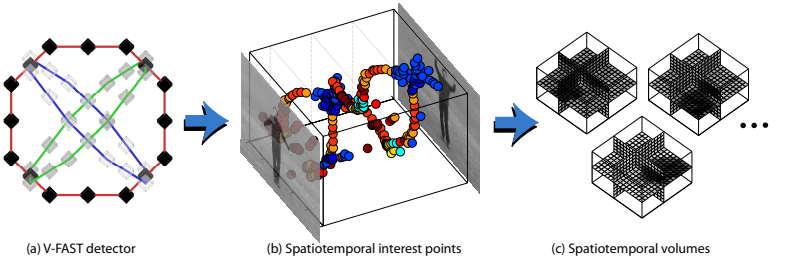


Figure 2: Spatiotemporal interest points localised by the proposed V-FAST detector

# 5 Spatiotemporal Semantic Texton Forests

Semantic texton forests [25] are ensembles of randomised decision trees which textonise input video patches into semantic textons. They are extremely fast to evaluate, since only a

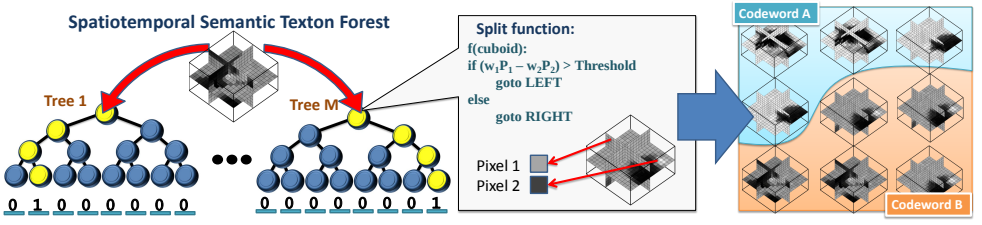


Figure 3: Visual codeword generation by Spatiotemporal Semantic Texton Forests

| Algorithm | Complexity | Relative Speed* | Hierarchical |
|---|---|---|---|
| k-means | $O(K)$ | 1 | no |
| Hierarchical k-means | $O(b\log_b(K))$ | 43.51 | yes |
| **STFs** | $O(\log_2(K))$ | 559.86 | **yes** |

*Speed measurements are relative to the k-means clustering algorithm. The speed is measured by computing 1 million feature vectors of 405 dimension. The codebook size $K$ is 1905 and the branching factor $b$ in the k-means algorithm is 16.

Table 1: A comparison of semantic texton forests and k-means codebooks.

small number of simple features are used to traverse the trees. They also serve a powerful discriminative codebook by multiple decision trees. Figure 3 illustrates how visual codewords are generated using the spatiotemporal semantic texton forests in the proposed method. It acts on small spatiotemporal volumes $p(x,y,t)$, which are taken around the detected interest points in input videos. The training process of STFs is similar to that of random forests. At each split node, candidate split functions are generated randomly, and the one that maximises the information gain ratio is chosen. The split functions in this work are defined as the weighted differences of two pixel values of the spatiotemporal volumes:

$$f(p) = w_1 \cdot p(x_1, y_2, t_1) - w_2 \cdot p(x_2, y_2, t_2) > threshold \tag{1}$$

The small volumes are passed down $M$ trees. The STF codebook has a size of $L = \sum_m^M L_m$ where $L_m$ represents the number of leaf nodes i.e. codewords in $m$-th tree. Figure 3 (right) shows the two codewords generated by the example split function. Table 1 summarises a comparison between STFs and k-means algorithms.

# 6 Pyramidal Spatiotemporal Relationship Match

Pyramidal spatiotemporal relationship match (PSRM) is presented to encapsulate both local appearance and structural information efficiently. Semantic texton forests quantise local space-time volumes into codewords in multiple texton trees. For each tree, the three-dimensional histogram is constructed by analysing pairs of codewords and their structural relations (see figure 4 (left and middle)). For each histogram, a novel pyramid match kernel is proposed for robust matching (figure 4 (right)). Multiple pyramidal matches are then combined to classify a query video. Whereas the spatiotemporal relationship match (SRM) [19] relies on a single flat k-means codebook, PSRM leverages the properties of semantic trees and pyramidal match kernels. Its hierarchical structure offers a time efficient way to perform the pyramid match kernel for semantic codeword matching [4].

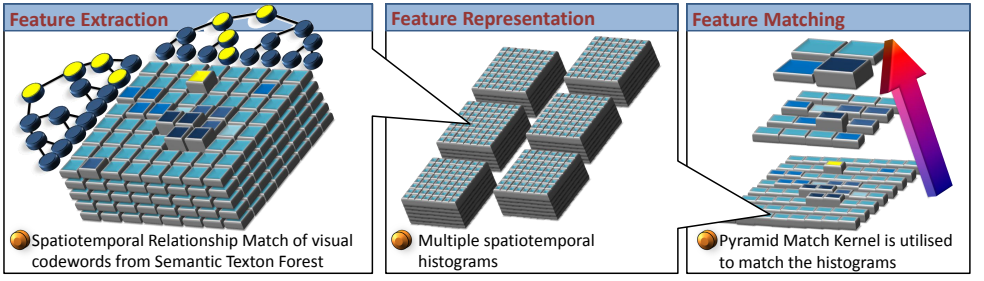


Figure 4: Pyramidal spatiotemporal relationship match (PSRM)

**Spatiotemporal relationship histograms.** Subsequences are sequentially sampled from an input video in very short intervals (*e.g.* 10 frames). A set of spatiotemporal interest points $\mathrm{U} = \{u_i\}$ are localised. The trained STFs assign visual codewords to the interest points. Therefore, an encoded interest point can be described as $u_i = \{x_i, y_i, t_i, l_{m,i}\}, m = 1, ..., M$, where $x_i, y_i, z_i$ represents a $XYT$-location of the feature and $l_{m,i}$ the visual codeword i.e. the leaf node assigned to $u_i$ by the $m$-th tree. A set of pairwise spatiotemporal associations are designed to capture the structural relations among interest points. By analysing all possible pairs $u_i$ and $u_j$ in U, space-time correlations are described by the following seven association rules $\mathrm{R} = \{R_1, \ldots, R_7\}$:

$$R_1\ overlap : |t_i - t_j| < \mathrm{T_o}, \qquad R_4\ nearXY : (|x_i - x_j| < \mathrm{T_n}) \wedge (|y_i - y_j| < \mathrm{T_n})$$
$$R_2\ before : \mathrm{T_o} < t_j - t_i < \mathrm{T_b}, \qquad R_5\ nearX : (|x_i - x_j| < \mathrm{T_n}) \wedge \sim (nearXY)$$
$$R_3\ after : \mathrm{T_o} < t_i - t_j < \mathrm{T_a}, \qquad R_6\ nearY : (|y_i - y_j| < \mathrm{T_n}) \wedge \sim (nearXY)$$
$$R_7\ far : (|x_i - x_j| < \mathrm{T_f}) \wedge (|y_i - y_j| < \mathrm{T_f}) \quad \wedge \sim (nearXY \vee nearX \vee nearY)$$

Figure 4 illustrates how the relationship histograms are constructed and matched using PSRM. A set of 3D relationship histograms $\{\mathrm{H}_1(\mathrm{U}), \ldots, \mathrm{H}_M(\mathrm{U})\}$ are constructed by analysing every pair of feature points in U. The bin $\mathrm{h}_m(i,j,k)$ of the $m$-th tree histogram $\mathrm{H}_m(\mathrm{U})$ takes the count of matching $(l_{m,i}, l_{m,j})$ codeword pairs by an association $R_k$. The total number of bins in $\mathrm{H}_m(\mathrm{U})$ is $L_m \times L_m \times |\mathrm{R}|$. Despite the large size of the relationship histograms, operations on these histograms can be greatly accelerated by sparse matrices.

**Pyramid match kernel for PSRM.** Similarity between the two sets of interest points U and V is measured by the pyramid match kernel (PMK) from a multi-resolution histogram space for each tree. At a specific resolution $q$, the two sets U and V having the histogram bins $\mathrm{h}_m^q(i,j,k)$ and $\mathrm{g}_m^q(i,j,k)$ respectively, are matched by histogram intersection in (2). New quantisation levels in the histogram pyramid are formed by increasing the bin size. In the proposed method, adjacent bins that share the same parent node in the tree are conveniently merged in (3), creating a new quantisation level $\mathrm{h}_m^{q+1}(i,j,k)$ (the same for $\mathrm{g}_m^{q+1}(i,j,k)$). The match kernel $\mathrm{K}_m$ at the $m$-th tree is then defined in (4) by the weighted summation of differences between successive histogram intersections. Matches in finer bins score higher

similarity than matches in coarser levels by a factor of $\frac{1}{4^{q-1}}$.

$$\mathrm{I}^q(\mathrm{U},\mathrm{V}) = \Sigma_{i=1}^{L_m}\Sigma_{j=i+1}^{L_m}\Sigma_{k=1}^{7}\left(\min(\mathrm{h}_m^q(i,j,k),\mathrm{g}_m^q(i,j,k))\right) \quad (2)$$

$$\mathrm{h}_m^{q+1}(i,j,k) = \Sigma_{u=1}^{2}\Sigma_{v=1}^{2}\left(\mathrm{h}_m^q(2(i-1)+u,2(j-1)+v,k)\right) \quad (3)$$

$$\mathrm{K}_m(\mathrm{U},\mathrm{V}) = \Sigma_{q=1}^{Q}\frac{1}{4^{q-1}}\left(\mathrm{I}^{q+1}(\mathrm{U},\mathrm{V})-\mathrm{I}^q(\mathrm{U},\mathrm{V})\right) \quad (4)$$

**Kernel k-means forest classifier.** We learn the k-means forest classifier using PSRM as a matching kernel. Given a set of training video data $\mathrm{U}_i$, $M$ independent clustering trees are grown by recursively performing k-means clustering on the pyramid matches. For the $m$-th tree in STFs, the hierarchical k-means algorithm aims to partition the training data into $\mathbf{S} = \{S_i\}, i = 1,...,N$ clusters so as to maximise the intra-cluster similarity by (5):

$$\arg\max_{\mathbf{S}} \Sigma_{i=1}^{N}\Sigma_{\mathrm{U}_j\in S_i}\mathrm{K}_m(\mathrm{U}_j,\mu_{m,i}) \quad (5)$$

where $\mu_{m,i}$ is the centroid of $i$-th cluster. In the testing stage, PSRM is performed on a query video V against all centroids $\mu_{m,i}$ at the same level. The query video proceeds to the node with the highest similarity score and PSRM is performed recursively until a leaf node is reached. Classification is done by the posterior probability by averaging the class distributions of the assigned leaf nodes $\{\hat{\mu}_m\}, m = 1,...,M$ trees as

$$\arg\max_c \mathrm{P}_H(c|\mathrm{V}) = \frac{1}{M}\Sigma_{m=1}^{M}\mathrm{P}_H(c|\hat{\mu}_m) \quad (6)$$

# 7 Combined Classification

**Bag of semantic textons.** The method called bag of semantic textons (BOST) developed for image classification [25] is applied to analyse local space-time appearance. The 1-D histogram B is obtained by counting the occurrences of interest points at every node in the STF codebook, hence the histogram size $|\mathrm{B}|$ is the total number nodes in the STFs. Since its dimension $L$ is relatively low (c.f. the PSRM histogram has $L_m \times L_m \times |\mathrm{R}|$ dimension), standard random forests [1] are applicable as a fast and powerful discriminative classifier, which is a proven technique in image categorisation and visual tracking. The random forests trained on the BOST histograms classify a query video V by the posterior probability by averaging the class distributions over the assigned leaf nodes $\{\hat{l}_1,\ldots,\hat{l}_m\}, m = 1,...M$ trees in the STFs: $\mathrm{P}_B(c|V) = \frac{1}{M}\sum_{m=1}^{M}\mathrm{P}_B(c|\hat{l}_m)$.

**Combined classification.** The task of action recognition is performed separately by the proposed kernel k-means forest classifier and by the BOST method. While PSRM has shown effective in most of the cases owing to its both local and structural information, BOST distinguishes classes that are structurally alike (*e.g.* walking and running). By integrating classification results of both methods, average accuracy is significantly improved. Final class labels are assigned to the classes $c$ which obtain the highest combined posterior probability as

$$\arg\max_c \mathrm{P}(c|\mathrm{V}) = \alpha_c\mathrm{P}_H(c|\mathrm{V}) + (1-\alpha_c)P_B(c|\mathrm{V}) \quad (7)$$

where the weight $\alpha_c$ is set to maximise the true positive ratio (sensitivity) of a class $c \in C$ by a gradient descent or line search.


Figure 5: Example frames of KTH (top row) and UT-interaction (bottom row) data sets

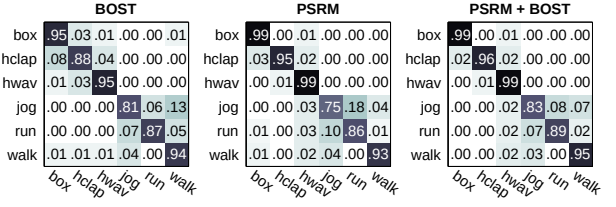


Figure 6: Confusion matrices of BOST (left), PSRM (middle), and combined classification(right) on KTH dataset

# 8 Experiments

The proposed method is tested on two public benchmarks, the KTH data set [23] and the UT-interaction data [20], a more challenging set [19]. Other published methods are compared with the proposed method in terms of recognition accuracy. Computational time of our method is also reported. Our prototype implemented by C++ in an Intel Core$^{TM}$ i7 920 PC showed real-time continuous action recognition performance.

## 8.1 KTH

The KTH data set, a common benchmark for action recognition research, involves sequences of six action classes taken with camera motions, scale, appearance and subject variations (see figure 5 (top)). To demonstrate the method for continuous action recognition by a short response time, subsequences of the length less than 2 seconds were extracted on the fly from the original sequences. The subsequences of training videos were used to build the classifiers. Similar subsequences were extracted from testing videos for evaluation. We used leave-one-out cross validation. Most published results in the literature were reported at the sequence level: class labels were assigned to whole testing videos instead of individual short subsequences. To put the proposed method in context, two different accuracies are measured: (1) the "snippet" accuracy that is directly measured at the subsequences level; and (2) the sequence level accuracy, which is measured by majority voting from the subsequences' classification labels.

Table 2 presents a detailed comparison of accuracies for our method and state-of-the-art methods. The PSRM+BOST model gives a very competitive accuracy despite that only short subsequences are used for recognition. The confusion matrices in figure 6 show how PSRM and BOST complement each other to attain an optimised accuracy. Quantisation effects are soothed by the multi-tree characteristics and pyramid matching of the proposed method, compared to the original spatiotemporal relationship match method [19].

Table 3 summarises the experiment results on recognition speed. Different from other sequence-level recognition approaches, a more realistic metric is designed to measure the al-

| Method | box | hclp | hwav | jog | run | walk | **Overall** | Protocol |
|---|---|---|---|---|---|---|---|---|
| **PSRM + BOST** | **100.0** | **96.0** | **100** | **86.0** | **95.0** | **97.0** | **95.67** | sequence |
| **PSRM + BOST** | **99.0** | **96.6** | **98.9** | **82.6** | **89.5** | **94.8** | **93.55** | snippet |
| PSRM | 99.0 | 96.1 | 98.7 | 74.6 | 85.9 | 92.2 | **91.10** | snippet |
| BOST | 94.8 | 88.2 | 95.0 | 81.3 | 87.2 | 94.0 | **90.10** | snippet |
| SRM [19] | 96.0 | 95.0 | 97.0 | 78.0 | 85.0 | 92.0 | **90.5** | sequence |
| Mined features (2009) [8] | 100.0 | 94.0 | 99.0 | 91.0 | 89.0 | 94.0 | **96.70** | sequence |
| CCA (2007) [5] | 98.0 | 100.0 | 97.0 | 90.0 | 88.0 | 99.0 | **95.33** | sequence |
| Neighbourhood** (2010) [6] | - | - | - | - | - | - | **94.53** | sequence |
| Info. maximisation (2008) [9] | 98.0 | 94.9 | 96.0 | 89.0 | 87.0 | 100.0 | **94.15** | sequence |
| Shape-motion tree (2009) [3] | 96.0 | 99.0 | 96.0 | 91.0 | 85.0 | 93.0 | **93.43** | sequence |
| Vocabulary forests (2008) [17] | 97.0 | 96.0 | 98.0 | 88.0 | 93.0 | 87.0 | **93.17** | sequence |
| Point clouds (2009) [11] | 95.0 | 93.0 | 99.0 | 85.0 | 89.0 | 98.0 | **93.17** | sequence |
| pLSA-ISM (2007) [23] | 96.0 | 92.0 | 83.0 | 79.0 | 54.0 | 100.0 | **83.92** | sequence |

* The length of subsequences called snippets is about 50 frames. To balance accuracy, speed and generality, the depth of random forest classifier = 8; For k-means forest classifier: K = 10, depth = 3. ** Classifiers were trained by a split dataset in separate scenarios.

Table 2: Accuracies on KTH data set by the proposed method and state-of-the-art methods. Leave-one-out cross validation (LOOCV) scheme was used.

| Dataset | V-FAST feature detection | STFs and BOST | PSRM | Random forests | k-means forests | **Total FPS** |
|---|---|---|---|---|---|---|
| KTH | 66.1 | 59.3 | 194.17 | 1137.6 | 67.1 | **18.98** |
| UT-interaction | 35.1 | 25.8 | 35.1 | 612.2 | 428.1 | **10.02** |

Table 3: Average recognition speed at different stages in frames per second (FPS)

gorithm speed. Every stage of the method (including feature detection, feature extraction and classification) is timed, and the average speed is defined as (*total number of subsequences*)/(*total recognition time*) **FPS**. It shows that the proposed method runs at 10 to 20 frames per second. The introduction of STFs has greatly improved the speed for feature extraction and codeword generation, outperforming the k-means visual codebook (see also Table 1). Using random forests and kernel k-means forest classifiers has provided a faster solution to match and classify multi-dimensional histograms over the traditional nearest neighbour and SVM classifiers.

## 8.2 UT-interaction data set

The UT-interaction data set contains six classes of realistic human-human interactions, including shaking hands, pointing, hugging, pushing, kicking and punching (see figure 5 (bottom)). Some challenging factors of this data set include moving backgrounds, cluttered scenes, camera jitters/zooms and different clothes. In the experiments, the segmented UT-interaction sequences were used for evaluating the recognition accuracy and speed of our method. As reported in table 4, the proposed method marked the best accuracy in classifying the challenging realistic human-human interactions. Under the complex human interactions, PSRM using both local appearance and structural cues appeared to be more stable than

| Method | shake | hug | point | punch | kick | push | **Overall** | Protocol |
|---|---|---|---|---|---|---|---|---|
| **PSRM+BOST** | **100.0** | **65.0** | **100.0** | **85.0** | **75.0** | **75.0** | **83.33** | sequence |
| PSRM | 90.0 | 50.0 | 85.0 | 65.0 | 70.0 | 40.0 | **66.67** | sequence |
| BOST | 80.0 | 50.0 | 100.0 | 65.0 | 25.0 | 35.0 | **59.16** | sequence |
| *SRM [19] | 75.0 | 87.5 | 62.5 | 50.0 | 75.0 | 75.0 | **70.8** | sequence |

* Unsegmented videos were used in the experiments.

Table 4: Accuracies on UT-interaction dataset. Leave-one-out cross validation (LOOCV) scheme was used.

BOST that uses only local appearance. However, there still exist improvements in overall recognition accuracies by the combined approach. The method runs at high speed more than 10 frames per second from table 3. The recognition speed on this data set over KTH has dropped due to extra interest points from other moving objects in the scene.

# 9 Conclusions

This paper has presented a novel real-time solution for action recognition. Compared to existing methods, a major strength of our method is in run-time speed. Real-time performance is achieved by semantic texton forests which work on video pixels generating visual codewords in an extremely fast manner. PSRM is proposed to capture both spatiotemporal structures and local appearances of actions and reduce quantisation errors. Furthermore, a novel fast interest point detector and application of random forests and kernel k-means forest classifiers contribute to the acceleration of recognition speed. Experimental results show the comparable accuracies of the proposed method over state-of-the-arts. Future challenges include tackling more complex realistic human actions and partial occlusions, as well as performing continuous action detection in real-time.

# References

[1] L. Breiman. Random forests. *Machine Learning*, 2002.

[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, 2005.

[3] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2009.

[4] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2005.

[5] T. Kim, S. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[6] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[7] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2003.

[8] Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2009.

**Acknowledgements.** Tsz-Ho Yu is funded by the Croucher Foundation.

[9] J. Liu and M. Shah. Learning human actions via information maximization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004.

[11] S. Gong M. Bregonzio and T. Xiang. Recognising action as clouds of space-time interest points. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[12] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[13] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[14] J. C. Niebles, H. Wang, and L. Fei-fei. Unsupervised learning of human action categories using spatial-temporal words. In *British Machine Vision Conference (BMVC)*, 2006.

[15] O. Oshin, A. Gilbert, I. Illingworth, and R. Bowden. Action recognition using randomized ferns. In *IEEE International Conference on Copmuter Vision Workshop on Video-Oriented Object and Event Classification*, 2009.

[16] A. Patron-Perez and I. Reid. A probabilistic framework for recognizing similar actions using spatio-temporal features. In *British Machine Vision Conference (BMVC)*, 2007.

[17] H. Riemenschneider, M. Donoser, and H. Bischof. Bag of optical flow volumes for image sequence recognition. In *British Machine Vision Conference (BMVC)*, 2009.

[18] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision (ECCV)*, 2006.

[19] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of copmlex human activities. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2009.

[20] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.

[21] S. Savarese, A. Del Pozo, J. Niebles, and L. Fei-fei. Spatial-temporal correlations for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing (WMVC)*, 2008.

[22] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *International Conference on Pattern Recognition (ICPR)*, 2004.

[24] P. Scovanner, S. Ali, and M.Shah. A 3-dimensional sift descriptor and its application to action recognition. In *International conference on Multimedia (MM)*, 2007.

[25] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[26] D. Tran and A. Sorokin. Human activity recognition with metric learning. In *European Conference on Computer Vision (ECCV)*, 2008.

[27] G. WIllems, J. H. Becker, T. Tuytelaars, and L. Van Gool. Exemplar-based action recognition in video. In *British Machine Vision Conference (BMVC)*, 2009.

[28] S. Wong, T. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[29] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2009.

[30] Z. Zhang, Y. Hu, S. Chan, and L. Chia. Motion context: A new representation for human action recognition. In *European Conference on Computer Vision (ECCV)*, 2008.