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

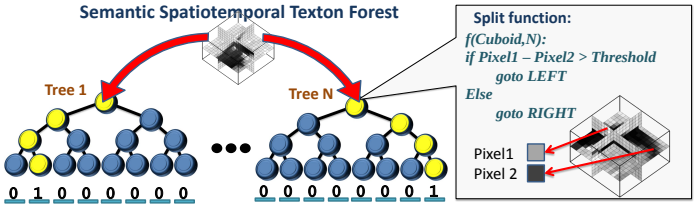

Figure 1: Visual codeword generation by Spatiotemporal Semantic Texton Forests

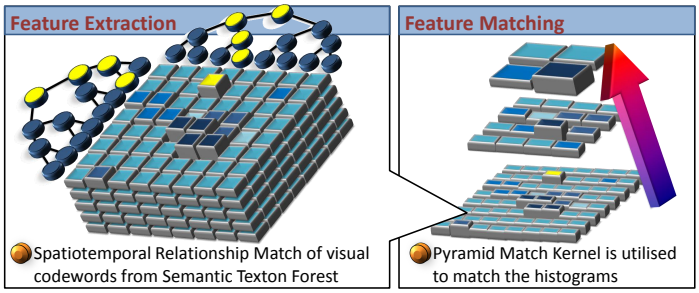

Figure 2: Pyramidal spatiotemporal relationship match (PSRM)

This paper presents a novel real-time action recogniser by utilising both local appearance and structural information. Our method is able to recognise actions continuously in real-time while achieving comparably high accuracy over state-of-the-arts.

Run-time speed is of vital importance in real-world action recognition systems, but existing methods seldom take computational complexity into full consideration. A class label is assigned after an entire query video is analysed, or a large lookahead is required to recognise an action. In addition, the "bag of words"(BOW) has proven effective for action recognition [5]. However, the standard BOW model ignores the spatiotemporal relationships among feature descriptors, which are useful for describing actions. Addressing these challenges, we present a novel approach for action recognition. The major contributions include the followings:

Efficient Spatiotemporal Codebook Learning: We extend the use of semantic texton forests [6] (STFs) from 2D image segmentation to spatiotemporal analysis. 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. STFs are ensembles of random decision trees that textonise input video patches into semantic textons. Since only a small number of simple features are used to traverse the trees, STFs are extremely fast to evaluate. They also serve a powerful discriminative codebook by multiple decision trees. Figure 1 illustrates how visual codewords are generated using STFs in the proposed method.

Combined Structural and Appearance Information: We propose a richer description of features, hence actions can be classified in very short video sequences. Based on [3], we introduce the pyramidal spatiotemporal relationship match (PSRM) to encapsulate both local appearance and structural information efficiently. Subsequences are sampled from an input video in short intervals (*e.g.* $\leq$ 10 frames). After spatiotemporal interest points are localised, the trained STFs assign visual codewords to the features. A set of pairwise spatiotemporal associations are designed to capture the structural relationships among features (*i.e.* pairwise distances along space-time axes). All possible pairs in the bag of features are analysed by the association rules and stored in the 3-D histogram. PSRM leverages the properties of semantic trees and pyramidal match kernels. Multiple pyramidal histograms are then combined to classify a query video. Figure 2 illustrates how the relationship histograms are constructed and matched using PSRM. For each tree in STFs, the three-dimensional histogram is constructed according to their spatiotemporal structures (see figure 2 (left)). Its hierarchical structure offers a time efficient way to perform the pyramid match kernel [1] for codeword matching (figure 2 (right)).

Enhanced Efficiency and Combined Classification: Several techniques are employed to improve the recognition speed and accuracy. A novel spatiotemporal interest point detector, called V-FAST, is designed based on the FAST 2D corners [2]. The recognition accuracy is enhanced by adaptively combining PSRM and the bag of semantic texton (BOST) method [6]: the k-means forest classifier is learned using PSRM as a matching kernel. 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, and BOST distinguishes classes that are structurally alike. By integrating classification results of both methods, average performance is significantly improved.

The proposed method is tested on two public benchmarks, the KTH data set [5] and the UT-interaction data set [4]. Other published results are compared with the proposed method in terms of recognition accuracy. Experimental results show the comparable accuracies of the proposed method over state-of-the-arts (figure 3). Furthermore, a major strength of our method over existing methods is in run-time speed. Real-time performance is achieved by semantic texton forest which work on video pixels generating visual codewords in an extremely fast manner.

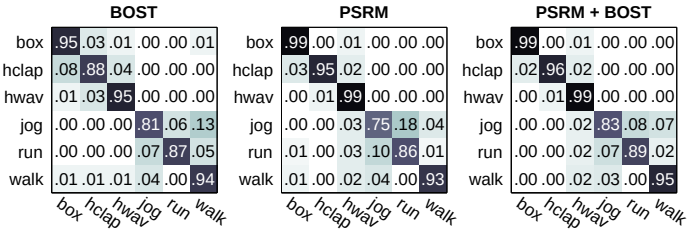

Figure 3: Confusion matrices of BOST (left), PSRM (middle), and combined classification(right) on KTH data set

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.