Real-time Activity Recognition by Discerning Qualitative Relationships Between Randomly Chosen Visual Features

Ardhendu Behera
http://www.comp.leeds.ac.uk/behera/
Anthony G Cohn
http://www.comp.leeds.ac.uk/agc/
David C Hogg
http://www.comp.leeds.ac.uk/dch/

School of Computing
University of Leeds
Leeds, LS2 9JT, UK
Email: {A.Behera, A.G.Cohn, D.C.Hogg}@leeds.ac.uk

Motivation. Automatic recognition of human *activities* (or *events*) from video is important to many potential applications of computer vision. One of the most common approach is the *bag-of-visual-features*, which aggregate space-time features globally, from the entire video clip containing complete execution of a single activity. The *bag-of-visual-features* does not encode the spatio-temporal structure in the video. For this reason, there is a growing interest in modeling spatio-temporal structure between visual features in order to improve the results of activity recognition.

The proposed framework. We model the spatio-temporal structure by exploiting the qualitative relationships between a pair of visual features. The proposed approach is inspired by [3, 4]. The goal is to find a pair of visual features whose spatiotemporal relationships are discriminative enough, and temporally consistent for distinguishing various activities. The framework is applied to recognize activities from a continuous live video (egocentric view) of a person performing manipulative tasks in an industrial setup. In such environments, the purpose of activity recognition is to assist users by providing on-the-fly instructions from an automatic system that maintains an understanding of the on-going activities.

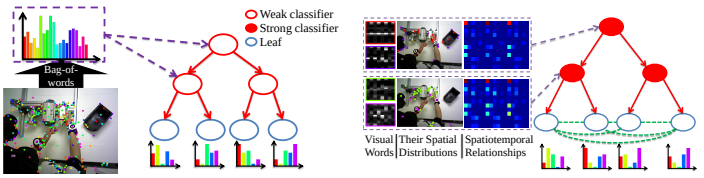
In order to recognize activities in real-time, we propose a *random forest with a discriminative Markov decision tree* algorithm that considers a random subset of relational features at a time and Markov temporal structure that provides temporally smoothed output (Fig. 1). Our algorithm is different from conventional decision trees [2] and uses a linear SVM as a classifier at each nonterminal node and effectively explores temporal dependency at terminal nodes of the trees. We explicitly model the spatial relationships of *left*, *right*, *top*, *bottom*, *very-near*, *near*, *far* and *very-far* as well as temporal relationships of *during*, *before* and *after* between a pair of visual features (Fig. 2), which are selected randomly at the non-terminal nodes of a given Markov decision tree. Our hypothesis is that the proposed relationships are particularly suitable for detecting complex non-periodic manipulative tasks and can easily be applied to the existing visual descriptors such as SIFT, STIP, CUBOID and SURF.

Growing discriminative Markov decision trees. Each tree is trained separately on a random subset of frames belonging to training videos. Learning proceeds recursively by splitting the training frames at internal nodes into the respective left and right subsets. This is done in the following four stages: randomly assign all frames from each activity class to a binary label; randomly sample a pair of visual words; compute the spatiotemporal relationships histogram $\mathbf{h}$ between them; and use a linear SVM to learn a binary split using the extracted $\mathbf{h}$. The binary SVM at each internal node sends the frame to the left child if $\mathbf{w}^T\mathbf{h} \leq 0$ otherwise to the right child, where $\mathbf{w}$ is the set of weights learned through the linear SVM. Using an information gain criteria, each binary split corresponds to a pair of visual words is evaluated on the training frames that falls in the current node. Finally, the split that maximizes the information gain is selected. The splitting process is repeated with the newly formed subsets until the current node is considered as a leaf node.

Inference. For real-time activity recognition, the proposed inference algorithm computes the posterior marginals $P(a_t|l_1^\tau \dots l_t^\tau)$ of all activities a_t over a frame I_t given a history of visited leaf nodes is $l_1^\tau \dots l_t^\tau$ (Fig.1b) for a particular tree τ. The smoothed output over the whole forest is achieved by averaging the posterior probabilities from all $\mathcal{T}$ trees:

$$a_t^* = \arg\max_{a_t} \sum_{\tau=1}^{\mathcal{T}} P(a_t|l_1^\tau \dots l_t^\tau)$$

Results. We evaluate our framework using an egocentric paradigm for recognizing complex manipulative tasks of assembling parts of a pump system in an industrial environment[1]. We compare our approach with our previous work in [1] which models the wrist-object and object-object interactions using qualitative and functional relationships. The accuracy of the proposed approach is 68.56% (using SIFT and STIP) and better than the method in [1], which is 52.09%. We also evaluated using *bag-of-visual-features* approach and the performance is 63.19%. This is achieved using a χ^2-SVM by concatenating STIP and SIFT *bag-of-visual-features*. Activity-wise performance comparison of live recognition is presented in Fig. 3.



(a) Conventional random DT (b) Discriminative Markov DT

Figure 1: (a) Conventional random Decision Trees (DT). The histogram below the leaf nodes represents the posterior probability distribution $P(a|l^\tau)$. (b) The proposed Markov DT sample a pair of visual words and the splitting criterion is based on the relationships between the sampled words. Green dotted lines illustrate the temporal dependencies.

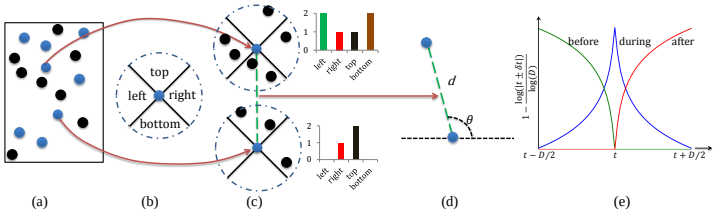


Figure 2: (a) A pair of visual word ('blue dots' and 'black dots') in an image. (b) *Local relationships* (c) Histogram representing *local relationships*. (d) *Global relationships* encode the oriented *very-near*, *near*, *far* and *very-far* relationships. (e) Temporal relationships of *before*, *during* and *after* over a sliding window of duration D.

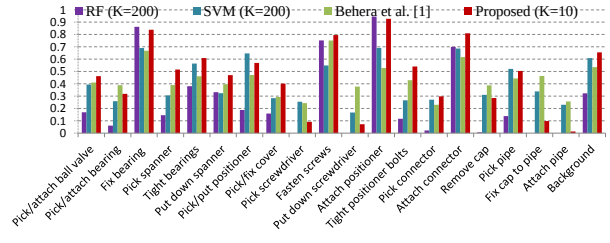


Figure 3: (a) Comparison of the performance of live activity recognition. SIFT bag-of-words ($K = 200$) results in accuracy of 53.21% using χ^2-SVM and 53.28% using conventional random forest. The method in [1] results in 52.09%. The proposed method is 66.20% ($K = 10$) significantly better than the baselines, where the random chance is 5%.

[1] A. Behera, D. C. Hogg, and A. G. Cohn. Egocentric activity monitoring and recovery. In *ACCV*, 2012.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

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

[4] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.

[1]Dataset and source code are available at `www.engineering.leeds.ac.uk/computing/research/ai/BallValve/index.htm`