Discriminative Topics Modelling for Action Feature Selection and Recognition

Matteo Bregonzio, Jian Li, Shaogang Gong, Tao Xiang
{bregonzio,jianli,sgg,txiang}@dcs.qmul.ac.uk

School of EECS, Queen Mary University of London
London, E1 4NS, U.K.

Problem - This paper addresses the problem of recognising realistic human actions captured in unconstrained environments (Fig. 1). Existing approaches for action recognition have been focused on improving visual feature representation using either spatio-temporal interest points or key-points trajectories. However, these methods are insufficient to handle the situations when action videos are recorded in unconstrained environments because: (1) Reliable visual features are hard to be extracted due to occlusions, illumination change, scale variation and background clutters. (2) Effectiveness of visual features are strongly dependent on the unpredictable characteristics of camera movements. (3) Complicated visual actions result in unequal discriminativeness of visual features.

Our Solutions - In this paper, we present a novel framework for recognising realistic human actions in unconstrained environments. The novelties of our work lie in three aspects: First, we propose a new action representation based on computing a rich set of descriptors from key point trajectories. Second, in order to cope with drastic changes in motion characteristics with and without camera movements, we develop an adaptive feature fusion method to combine different local motion descriptors for improving model robustness against feature noise and background clutters. Finally, we propose a novel Multi-Class Delta Latent Dirichlet Allocation (MC-ΔLDA) model for feature selection. The most informative features in a high dimensional feature space are selected collaboratively rather than independently.

Motion Descriptors - We first compute trajectories of key-points using KLT tracker and SIFT matching. After trajectory pruning by identifying the Region of Interest (ROI), we compute three types of motion descriptors from the survived trajectories. First, Orientation-Magnitude Descriptor is extracted by quantising orientation and magnitude of motion between two consecutive points in the same trajectory. Second, Trajectory Shape Descriptor is extracted by computing Fourier coefficients of a single trajectory. Finally, Appearance Descriptor is extracted by computing the SIFT features at all points of a trajectory.

Interest Point Features - We also detect spatio-temporal interest points as they contain complementary information to trajectory features. At an interest point, a surrounding 3D cuboid is extracted. We use gradient vectors to describe these cuboids and PCA to reduce descriptor's dimensionality.

Adaptive Feature Fusion - We wish to fuse adaptively trajectory based descriptors with 3D interest point based descriptors according the presence of camera movement. The presence of moving camera is detected by computing the global optical flow over all frames in a clip. If the majority of the frames contain global motion, we regard the clip as being recorded by a moving camera. For clips without camera movement, both interest point and trajectory based descriptors can be computed reliably and thus both types of descriptors are used for recognition. In contrast, when camera motion can be detected, interest point based descriptors are less meaningful so only trajectory descriptors are employed.

Collaborative Feature Selection - We propose a MC-ΔLDA model (Fig. 2) for collaboratively selecting dominant features for classification. We consider each video clip $\mathbf{x}_j$ is a mixture of N_t topics $\Phi = \{\phi_t\}_{t=1}^{N_t}$ (to be discovered), each of which ϕ_t is a multinomial distribution over N_w words (visual features). The MC-ΔLDA model aims to constrain topic proportion *non-uniformly* and on a per-clip basis. For each video clip belonging to action category A_c, we model it as a mixture of: (1) N_t^s topics which are shared by all N_c category of actions, and (2) $N_{t,c}$ topics which are uniquely associated with action category A_c. In MC-ΔLDA, the non-uniform proportion of topic mixture for a single clip $\mathbf{x}_j$ is enforced by its action class label c_j and the hyperparameter α^c for the corresponding action class c. Given the total number of topics $N_t = N_t^s + \sum_{c=1}^{N_c} N_{t,c}$, the structure of the MC-ΔLDA model, and the observable variables (clips $\mathbf{x}_j$ and action labels c_j), we can learn the N_t^s shared topics as well as all $\sum_{c=1}^{N_c} N_{t,c}$ unique topics for all N_c classes of actions. We use the N_t^s topics shared by all actions for selecting discriminative features. The N_t^s shared topics are represented as an $N_w \times N_t^s$ dimension matrix $\hat{\Phi}^s$. The feature selection can be summarised into two steps: (1) For each feature v_k, $k = 1, \cdots, N_w$, compute its maximum probability across all N_t^s topics according to $p(v_k) = \max(\hat{\Phi}^s_{k,1:N_t^s})$; (2) Rank the value of $p(v_k), k = 1, \cdots, N_w$ in ascending order to obtain a vector of feature index r(V) in which higher ranked features correspond to more discriminative/relevant features.

Figure 1: Actions captured in an unconstrained environments, YouTube dataset. From left to right: cycling, diving, soccer juggling, and walking with a dog.

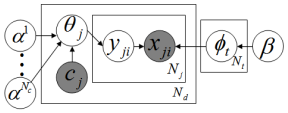

Figure 2: MC-ΔLDA model.

	UCF Films	UCF Sports	YouTube
Our model	**96.75%**	86.90%	64.00%
Wang et al. [4]	86.60%	85.60%	-
Yeffet et al. [6]	80.75%	79.20%	-
Rodriguez et al. [3]	66.30%	69.20%	-
Kovashka et al. [1]	-	**87.20%**	-
Yao et al. [5]	-	86.60%	-
Liu et al. [2]	-	-	**71.20%**

Table 1: Comparison with the state-of-the-art.

	UCF Films	UCF Sports	YouTube
MC-ΔLDA	**96.75%**	**86.90%**	**64.00%**
Mutual Information [7]	96.10%	85.33%	62.20%
No Feature Selection	96.10%	84.00%	59.90%

Table 2: Comparing effectivenss of different feature Selection methods.

Results - Three action datasets were used in our experiments: **UCF Feature Films Dataset** [3] providing a representative pool of natural samples of two action classes including Kissing and Hitting/slapping, **UCF Sport Actions Dataset** [3] containing 10 different types of human actions in sport broadcasting videos, and **YouTube Dataset** [2] which is the most extensive realistic dataset available composed of 1168 videos collected from YouTube. The average recognition rates obtained are compared with the existing approaches in Table 1. We achieved significant improvement on the UCF Films dataset and comparable results on the UCF Sports. On the YouTube Dataset the performance is slightly worse than that in [2] which uses quite different features and classifiers, and relies on a number of heuristic preprocessing steps that are hard to reproduce. Table 2 shows that our collaborative feature selection method improves the action recognition performance and ourperform a mutual information based sequential feature selection method.

[1] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.

[2] J. Liu and J. Luo snd M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

[3] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[4] H. Wang, M. Muneeb Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[5] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *CVPR*, 2010.

[6] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009.

[7] M. Zaffalon and M. Hutter. Robust feature selection by mutual information distributions. In *UAI*, pages 577–584, 2002.