Discriminative Topics Modelling for Action Feature Selection and Recognition

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

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

Abstract

This paper presents a framework for recognising realistic human actions captured from unconstrained environments. The novelties of this 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 model for feature selection. The most informative features in a high dimensional feature space are selected collaboratively, rather than independently as by existing feature selection methods. Extensive experiments on challenging public datasets demonstrate the effectiveness of the proposed framework.

1 Introduction

Recent interest in human action recognition research has been shifted from recognising actions captured by a single staged action per video clip in a well-controlled environment [2, 16], to more realistic actions captured from an unconstrained environment in feature films, sports broadcasting videos [15] and home videos on YouTube [10] (Fig. 1). Action recognition in an unconstrained environment, also referred to as "in the wild" [10], is challenging due to a number of reasons. First, action videos in the wild are subject to a much greater degree of occlusions from multiple objects, illumination change, shadow, cluttered background, scale variation, low spatial and temporal video resolution. They pose serious problems in feature extraction for action representation as one must compute and select the more informative and relevant visual features for each action of interest, and igore features caused by background clutter and other moving objects in the scene. Second, cameras can be either static or moving in an unpredictable manner. With camera movement, motion features are contributed by both the action and cluttered background. This can affect the usefulness

BMVC 2010 doi:10.5244/C.24.8

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

of extracted features for action representation as feature effectiveness is dependent on the characteristics of camera movements. Finally, given an action representation with a large pool of features, not all features are equally informative and discriminative. Feature selection is required. However, good feature selection in a high dimensional space is hard when different features from both foreground and background are closely correlated.

In this paper, we propose a novel framework to address these three problems for recognising actions in the wild. Our framework three novel features:

A rich motion descriptor for action representation. Local image features have been widely used for action recognition. These features can be extracted from spatio-temporal 3D (2D+time) interest points [4, 6, 9, 10, 14] or key-points trajectories [5, 7, 12, 17]. 3D interest points correspond to local spatio-temporal saliency of an image sequence from which both local motion and appearance characteristics can be extracted. However, features from 3D interest points are limited in temporal scalability as they only capture short (simple) movements within a short temporal window therefore are inadequate for describing longer-term and more complex movements. Alternatively, long-term motion characteristics can be extracted from trajectories through tracking key points. However, this approach is both vulnerable to low-textured environments (not enough key points) and does not retain static appearance information, which has a role to play in recognising actions of subtle movements. In this paper, we formulate a novel trajectory representation which is able to retain local motion information (trajectory orientation and magnitude), trajectory shape and static appearance information. Our descriptors are invariant to changes in scale, action direction, frame resolution and provide more robust description comparing to existing method [17].

Adaptive feature fusion. We consider both key point trajectory based descriptors and 3D interest point based descriptors since they are complementary. However, camera movements have very different effects on these two types of descriptors. In particular, without camera movement, both descriptors can be extracted reliably. However, given unpredictable camera movements, 3D interest point based descriptors become much more unreliable. To address this problem, we formulate an adaptive fusion method for selectively combining these two types of features.

Collaborative feature selection. To select more informative and discriminative features for action recognition, we propose a novel Multi-Class Delta Dirichlet Allocation (MC-ΔLDA) topic model for collaborative feature selection. Traditionally, feature selection has been performed through expensive sequential search in a large feature space [21] which mostly ignores any correlation between different features. The proposed MC-ΔLDA is designed to retain any correlation among features and select them collaboratively. MC-ΔLDA is an extension of ΔLDA proposed by [1], which was used for understanding code bugs in computer programs with binary document classes (with and without bugs). In this work, the

(a) (b) (c)

Figure 2: (a) Orientation-Magnitude Descriptor: An action trajectory is quantized and converted to a histogram. (b) All the detected trajectories. (c) Trajectories of interest and ROI.

formulated MC-ΔLDA is aimed at discovering action features groups (topics), some of them corresponding to features shared across different actions categories and others corresponding to unique features from specific action classes. By grouping all features jointly and collaboratively, MC-ΔLDA provides more effective feature selection for action discrimination.

Extensive experiments were conducted to evaluate the effectiveness of the proposed framework using realistic action datasets including the YouTube Dataset [10], the UCF Sports dataset and Feature Films dataset [15]. Our results demonstrate that the proposed methods significantly outperform existing techniques on the UCF Feature Films dataset whilst achieving comparable results on both YouTube and UCF Sports datasets.

2 Action Representation

2.1 Key Point Trajectory Features

Trajectory Generation – We first compute trajectories of key-points using two techniques: the Pyramid Lucas-Kanade-Tomasi (KLT) tracker [11, 12] and the SIFT matching [17]. These two trackers are applied independently to a video footage so that we can obtain trajectories as dense as possible even in low textured videos. During tracking, we consider a trajectory "reliable" if it lasts more than 5 frames. Any shorter trajectory is automatically removed. When a trajectory reaches a pre-defined max length (25 frames), it is auto-segmented and a new trajectory is created.

Trajectory Pruning – Extracted trajectories in a video may not always be useful for action recognition. For example, in Fig. 2 (b), lots of trajectories are extracted from the background area and thus need to be removed in order to retain the most relevant trajectories for describing the body actions of the person (Fig. 2 (c)). To that end, we consider a trajectory pruning process. Our approach is based on the detection a region of interest (ROI) from each video frame by measuring trajectory similarity within a temporal window. Suppose there exists N trajectories that pass through a frame f: $\mathbf{T} = \{\mathbf{t}_i\}, i = 1, \cdots, N$. For each trajectory, we define a trajectory segment $\mathbf{t}_i$ within a temporal window of 4 frames centred at frame f and each framewise displacement vector $\mathbf{d}_i$ as: $\mathbf{t}_i = \{(x^i_{f-1}, y^i_{f-1}), (x^i_f, y^i_f), (x^i_{f+1}, y^i_{f+1}), (x^i_{f+2}, y^i_{f+2})\}$, and $\mathbf{d}_i = \{d^i_1, d^i_2, d^i_3\}$, in which $d^i_k = \left(x^i_{f+(k-1)} - x^i_{f+(k-2)}, y^i_{f+(k-1)} - y^i_{f+(k-2)}\right)$. We aim to measure similarity between any pair of trajectory displacement vectors $\mathbf{d}_i$ and $\mathbf{d}_j$. This results in an $N \times N$ dimensional similarity matrix $\mathbf{C}$ with component:

$$\mathbf{C}_{i,j} = \sum_{k=1}^{3} \|d^i_k - d^j_k\|, \tag{1}$$

from which we compute a single similarity score for trajectory $\mathbf{t}_i$ as $m_i = \sum_{j=1}^{N} \mathbf{C}_{i,j}$. This score measures the similarity of this trajectory to all the other trajectories within a 4-frame

temporal window centred at the frame f. We consider that if a trajectory is very similar to the others, it is likely that it is extracted from background clutters thus should be removed. To this end, in each frame f, we compute an adaptive threshold $M_{TH}^f = \frac{\gamma}{N}\sum_{i=1}^{N} m_i$ where γ is empirically set to 1.3 in our experiment and remove any trajectories whose similarity score is larger than M_{TH}^f. After thresholding, assume N_f reliable trajectories are left in the frame f, we can now compute the centroid of the region of interest (ROI) by averaging spatial coordinates all key points on the reliable tracks:

$$\hat{x} = \frac{1}{N_f}\sum_{i=1}^{N_f} x_f^i, \qquad \hat{y} = \frac{1}{N_f}\sum_{i=1}^{N_f} y_f^i \tag{2}$$

and its dimensions are given by $D_x = 2\sqrt{2c_{xx}}$ and $D_y = 2\sqrt{2c_{yy}}$ where c_{xx} and c_{yy} are the second central moments of reliable key points. Any trajectory of interest which is located outside the ROI will then be removed. An example of remaining trajectories after pruning is shown in Fig. 2 (c). It shows clearly that the region of interest corresponds accurately to where the action takes place. Region of interest detection has been exploited elsewhere [10] based on 2D interest point detection. In contrast, our approach is based on statistical analysis of key point trajectory distributions and is robust for videos captured by both static and moving cameras.

Orientation-Magnitude Descriptor – Given two consecutive points: $P = (x_l, y_l)$, $P' = (x_{l+1}, y_{l+1})$, along a trajectory as illustrated in Fig. 2(a), we compute the displacement vector between the two points as $d = (x_{l+1} - x_l, y_{l+1} - y_l)$. For a single trajectory $\mathbf{t}$ of length L we can then obtain a series of displacement vectors $\mathbf{d} = \{d_1, d_2, ..., d_{L-1}\}$. We perform quantization on $\mathbf{d}$ by considering both magnitude and orientation of the displacement vectors. For magnitude quantization, we normalise each displacement vector by the largest displacement magnitude within the same trajectory and apply 4 uniform quantization levels. For orientation quantization, we divide top and bottom half circles into 8 equal sectors, each subtending 22.5°, as shown in Fig. 2(a). The formulated quantization results in a track descriptor that is both scale-invariant and direction-invariant. Combining magnitude and orientation quantizations, each trajectory is then described by a 32-bin histogram O.

Trajectory Shape Descriptor – Complex Fourier descriptors are commonly used to represent and compare shapes extracted from object silhouette for object recognition [22]. Here we formulate a Fourier descriptor to describe the shape signature of a trajectory. Assume a trajectory consists of L key points $\{(x_1, y_1), (x_2, y_2), ..., (x_{L-1}, y_{L-1},)\}$. We aim to describe this trajectory as a 2D shape of N vertices $\{z(i) : i = 1, \ldots, N\}$. These N vertices can be computed using the N coefficients of the Fourier transform of $\{z(i)\}$:

$$z_i = \sum_{k=\frac{-N}{2}+1}^{\frac{N}{2}} c_k \exp\left(2\pi j \frac{ki}{N}\right). \tag{3}$$

The Fourier coefficients c_k present the frequency contents of the trajectory in which lower frequency components describe approximative shape while higher frequency components retain more trajectory details. They provide an useful descriptor for trajectory global characteristics. Within the N Fourier coefficients, c_0 is omitted because it represents centre of gravity of a trajectory and by removing this term, the descriptor is invariant to translation. Moreover, we use c_1 to normalise all other Fourier coefficients, making them be invariant to homothety transformations. As a result, each trajectory is represented by an $N-1$ dimension vector F. Note that our Fourier descriptor is very different from the orientation-magnitude

descriptor in that the former is a global shape descriptor concerning global motion information along a trajectory whilst the latter is a bag-of-words (BOW) descriptor of local motion information of track segments. Both types of descriptors contain complimentary information.

Appearance Descriptor – Given a trajectory with a length L, we first extract SIFT features S_i at all the L key points $i = 1,...,L$. An appearance description of the tracked key point with this trajectory is computed as the average of L SIFT features: $S = \frac{1}{L}\sum_{i=1}^{L} S_i$.

Holistic Trajectory Representation – In order to fuse all three descriptors for action representation, we employ the BOW method. For each trajectory, its associated descriptors O, F and S are normalized and concatenated to form a global descriptor $G = [O, F, S]$. The creation of BOW representation of a video consists of two steps. First we quantise the global descriptors G for all trajectories using K-means to obtain a codebook with 500 words and assign each trajectory a codeword. Second, to retain spatio-temporal information about trajectories, we divide the spatio-temporal volume of a ROI in a video into 8 blocks including 4 non-overlapping spatial blocks and 2 temporal overlapping blocks (with a length of 2/3 of the temporal window of the ROI volume). Trajectory in each spatio-temporal blocks are then labelled separately. This results in a codebook V_1 with $500 \times 8 = 4000$ codewords to describe trajectories in videos.

2.2 Spatio-Temporal Interest Points Features

We also consider local features extracted based on spatio-temporal interest points as they contain complementary information to our trajectory features. In our model, 3D interest points are detected using the method of [4]. Compared with alternative methods such as [6], this approach is more reliable under realistic conditions such as small camera movement, camera zooming, and shadows. The interest points are selected at the local maximal of detector response, and 3D cuboids are extracted around them. Similar to [6, 10], we use gradient vectors to describe these cuboids and PCA to reduce the descriptor's dimensionality. To reduce the effect of spurious detection, we employ an outlier removal method which deletes points far from the mass centre of the points cloud. Bag-of-word is deployed again to represent each video clip. Specifically, we initially build a codebook V_2 with 300 visual words by performing k-means to a random subset local features from the training data. Then, we represent each clip with a visual-words histogram.

2.3 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. This simple method can accurately and consistently separate videos with and without camera movements. 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, resulting in a final codebook $V = [V1, V2]$ with 4300 visual words. In contrast, when camera motion can be detected, interest point based descriptors are less meaningful so only trajectory descriptors are employed, resulting the final codebook $V = V1$ with 4000 visual words. We now represent a set of N_d video clips as $\mathbf{X} = \{\mathbf{x}_j\}$, $j = 1, \cdots, N_d$, each of which is represented as a set of labelled features using V.

3 Collaborative Feature Selection

3.1 Multi-Class Delta Latent Dirichlet Allocation (MC-ΔLDA)

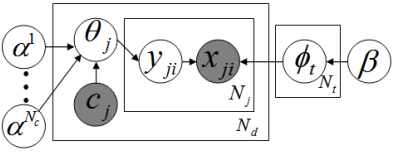


Figure 3: MC-ΔLDA model.

For MC-ΔLDA modelling, 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 multi-nominal distribution over N_w words (visual features). However, different from existing topic models such as LDA model [4] which assumes uniform proportion of topic mixture for each video clip, a MC-ΔLDA model aims to constrain topic proportion *non-uniformly* and on a per-clip basis. More precisely, 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. The structure of the proposed MC-ΔLDA model is shown in Fig. 3. 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 $N_t = N_t^s + \sum_{c=1}^{N_c} N_{t,c}$ and let T_0 be the first N_t^s elements list of shared topics and T_c be the $N_{t,c}$ element list of topics for action c, each hyperparameter α^c is a vector with N_t components $\alpha^c = \{\alpha_t^c\}_{t=1}^{N_t}$ in which components $t \in T_0 \bigcup T_c$ are constrained to be non-zero. To enforce non-uniform proportion of topic mixture, a generative process of sampling video clips is given as follows:

1. Draw a Dirichlet word-topic distribution $\phi_t \sim \text{Dir}(\beta)$ for every topic t;
2. For each document j:
   (a) Draw a class label $c_j \sim \text{Multi}(\varepsilon)$;
   (b) Given label c_j, draw a constrained topic distribution $\theta_j \sim \text{Dir}(\alpha^{c_j})$;
   (c) Draw a topic $y_{j,i}$ for each word i from multinomial $y_{j,i} \sim \text{Multi}(\theta_j)$;
   (d) Sample a word $x_{j,i}$ according to $x_{j,i} \sim \text{Multi}(\phi_{y_{j,i}})$.

Given the structure of the MC-ΔLDA model and observable variables (clips $\mathbf{x}_j$ and action labels c_j), our objective is to 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. The full joint probability of a document j in MC-ΔLDA is

$$p(\mathbf{x}_j, \mathbf{y}_j, \theta_j, \Phi, c | \alpha, \beta, \varepsilon) = \prod_i p(x_{j,i}|y_{j,i}, \Phi) p(y_{j,i}|\theta_j) p(\theta_j|\alpha, c) p(c|\varepsilon) p(\Phi|\beta). \quad (4)$$

Similar to the standard LDA, exact learning in our model is intractable. However a collapsed Gibbs sampler can be derived to sample the topic posterior $p(\mathbf{y}|\mathbf{x}, c, \alpha, \beta)$ (now additionally conditioned on the current class c) leading to the update

$$p(y_{j,i}| \quad \mathbf{y}_{j,-i}, \mathbf{x}, c, \alpha, \beta) \quad \propto \frac{n_{x,y}^{-i} + \beta}{\sum_x n_{x,y} + \beta} \frac{n_{y,d}^{-i} + \alpha_y^{c_j}}{\sum_y n_{y,d}^{-i} + \alpha_y^{c_j}}. \quad (5)$$

Here $\mathbf{y}_{j,-i}$ indicates all topics except the token i; $n_{x,y}^{-i}$ indicates the counts of topic y being assigned to word x, excluding the current item i; and $n_{y,d}^{-i}$ indicates the count of topic y

occurring in the current document d. These counts are also used to point estimate Dirichlet parameters θ and Φ by their mean, e.g.

$$\hat{\Phi}_{x,y} = \frac{n_{x,y}+\beta}{\sum_x n_{x,y}+\beta}. \quad (6)$$

In our model, we set the number of shared topics to $N_t^s = 5$ and assigned each action category a single unique topic $N_{t,c} = 1$. Moreover, we set the non-zero elements of α^c for shared topics to 0.1 and for unique topics to 10. The hyperparameter β is learned with Gibbs-expectation maximization [13].

3.2 Feature Selection using MC-ΔLDA

Using MC-ΔLDA enables us to learn N_t topics $\hat{\Phi}$ to represent natural grouping of visual features either shared by all classes of actions or uniquely associated with one particular action category.This grouping process effectively ranks all the visual features according to their importance of being used for representing either general aspects of all action classes or unique aspects of a specific action class. Now, we need to learn a sorted index of all the visual features $r(V)$ from the learned topics to represent distinctive visual features for action classification. Ideally, the ranked features $r(V)$ can be learned from the $\sum_{c=1}^{N_c} N_{t,c}$ action specific topics in $\hat{\Phi}$. However, as visual features extracted from action videos can be very noisy and not well structured, those topics can be easily corrupted by noise and are thus not suited for feature selection. Therefore we learn the discriminative features from the N_t^s topics shared by all actions. Given the N_t^s shared topics which 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.

As our model select different types of features to represent videos with and without camera movements, different MC-ΔLDA models are trained separately for the two type of videos. For each model, the number of features selected for final classification are determined by cross validation.

4 Experiments

Datasets – Three action datasets were used in our experiments. **UCF Feature Films Dataset** [15] provides a representative pool of natural samples of two action classes including Kissing and Hitting/slapping. It contains 92 samples of Kissing and 112 samples of Hitting/Slapping, extracted from a range of classic movies. The actions were captured in a wide range of scenes under different viewpoints with different camera movement patterns. The video clips have different frame rate and different image size, lasting between 5 to 15 seconds. **UCF Sport Actions Dataset** [15] contains 10 different types of human actions in sport broadcasting videos: diving, kicking , weight-lifting, horse-riding, running, skateboarding, golf swinging, swinging 1 (gymnastics, on the pommel horse and floor), swinging 2 (gymnastics, on the high and uneven bars) and walking. The dataset consists of 150 video samples which show a large intra-class variability. The videos have different frame rate and image size. They last in average 5 seconds. **YouTube Dataset** [10] is the most extensive realistic action

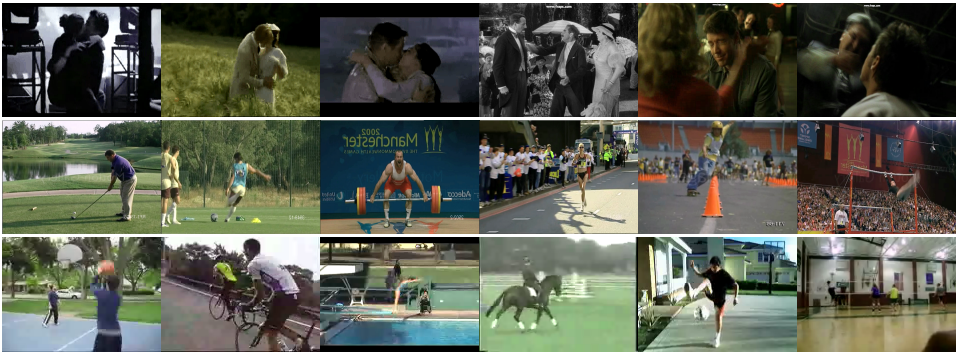
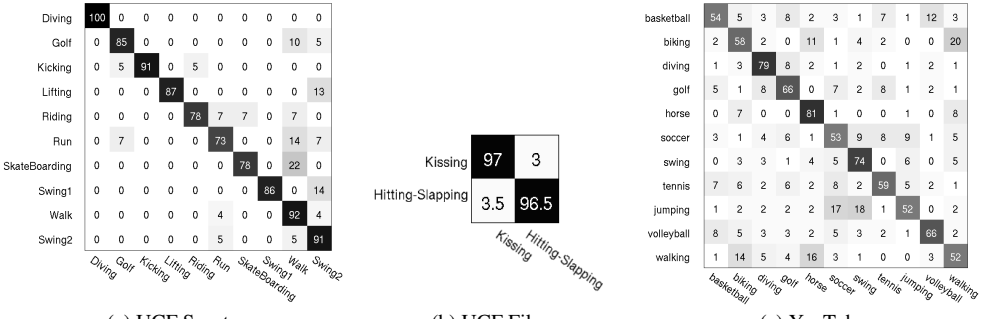


Figure 4: From top to bottom: Example frames from the UCF Feature Films, the UCF Sport Actions and the YouTube datasets.



(a) UCF Sports (b) UCF Films (c) YouTube

Figure 5: Confusion matrices of our approach on three datasets.

dataset available to public. it is composed of 1168 videos collected from YouTube. These videos contain a representative collection of real world challenges such as: shaky cameras, cluttered background, variation in object scale, variable and changing view-point and illumination, and low resolution. Particularly, since these videos are mostly home-videos captured by hand-held cameras, the camera movements are much more unpredictable compared the other two datasets. The YouTube dataset contains 11 action categories: basketball shooting, volleyball spiking, trampoline jumping, soccer juggling , horse-back riding, cycling, diving, swinging, golf swinging, tennis swinging, and walking. Clips have different frame rate by constant frame size of 320 by 240 pixels. The clips last in between 3 and 15 seconds. Examples of the three datasets are shown in Fig. 4.

Settings – Recognition was performed using Support Vector Machine with a Gaussian kernel. We used Leave-One-Out Cross-Validation (LOOCV). More specifically, for the YouTube dataset, adopting the settings given in [10], the dataset was divided into 25 subsets, out of which 24 subsets were used for training and the remaining subset was used for testing. For the UCF Sport Actions and Feature Films datasets we follow the setting in [15]: one clip was used for testing and the renaming for training. In our experiment, we empirically used 29 Fourier coefficients in the trajectory shape descriptor, and for the appearance descriptor, we used 128-bin SIFT histogram.

Comparison with state-of-the-art techniques – The average recognition rates obtained using the proposed approach are presented in Table 1 and compared with the results obtained by existing approaches. The classification confusion matrices are also presented in Fig. 5.

	UCF Films	UCF Sports	YouTube
Our model	**96.75%**	86.90%	64.00%
Wang et al. [18]	86.60%	85.60%	-
Yeffet et al. [20]	80.75%	79.20%	-
Rodriguez et al. [15]	66.30%	69.20%	-
Kovashka et al. [8]	-	**87.20%**	-
Yao et al. [19]	-	86.60%	-
Liu et al. [10]	-	-	**71.20%**

Table 1: Comparison on the UCF Feature Films, UCF sport actions and YouTube datasets.

	UCF Films	UCF Sport	YouTube
OM	80.50%	58.04%	44.7%
F	82.00%	75.02%	45.7%
S	91.60%	82.50%	44.5%
OM+F+S	94.40%	84.45%	59.90%
OM+F+S+IP without adaptive fusion	92.20%	77.33%	49.03%
OM+F+S+IP with adaptive fusion	96.75%	86.90%	64.00%

Table 2: Evaluation of descriptor performances and effectiveness of feature fusion. **OM**: Orientation-Magnitude Descriptor, **F**: Fourier Descriptor, **S**: SIFT Descriptor, and **IP**: spatio-temporal interest points.

Excellent result is obtained on the UCF Feature Film dataset with 96.75% average recognition rate for the two action classes. This result is significantly better than those obtained by existing approaches, best of which [18] achieved 86.60%. Our approach also outperforms all existing methods which report results on the UCF Sport dataset except [8]. As for the YouTube Dataset, an average recognition rate of 64.00% was obtained. This is a much harder dataset compared to the other two due to its home video nature. Apart from the work [10] which first introduced this dataset, no other work seems to be able to present results on this dataset. Our results is slightly lower than that in [10] which used quite different features and different classifiers, and include a number of heuristic steps that are hard to reproduce.

Effectiveness of Adaptive Feature Fusion – In Tab. 2 we show the effectiveness of each single descriptor and the fusion of them. It is evident when the three trajectory based descriptors are fused together, action recognition performance is improved for all three datasets. The improvement is particularly significant for the YouTube dataset (around 15% increase compared to any single descriptor alone), which contains a large variety of action categories and vastly different lighting conditions, camera angles, and camera movements. It is thus more important for different complimentary information to be utilised simultaneously. Table 2 also shows that fusion of trajectory based features with 3D interest point based features can lead to better performance, provided that they are fused in an adaptive manner as proposed in this paper. In particular, if they are fused unconditionally without considering the reliability of each type of feature given the camera movements, performance degradation is observed. This result validates the effectiveness of the novel action representation and feature fusion method formulated in this work.

Effectiveness of Collaborative Feature Selection – In Table 3 we compare the effectiveness of our collaborative feature selection method with a mutual information based sequential feature selection method proposed in [21]. It is evident that our feature selection method indeed improves action recognition performance compared to using all features without selection. The effect from our feature selection is in particular more significant for the more difficult YouTube dataset. In this dataset the unpredictable camera movements introduced large number of irrelevant features which cannot be removed completely even with our trajectory prun-

	UCF Films	UCF Sports	YouTube
MC-ΔLDA	**96.75**%	**86.90**%	**64.00**%
Mutual Information [21]	96.10%	85.33%	62.20%
No Feature Selection	96.10%	84.00%	59.90%

Table 3: Comparing effectivenss of different feature Selection methods.

ing and region of interest detection. In comparison, the sequential feature selection process is less effective. This suggests the importance of performing feature selection jointly and collaboratively given highly correlated features extracted both instantaneously from interest points and globally over time from trajectories.

5 Conclusions

We presented a novel framework for adaptive fusion and collaborative selection of visual features for recognising realistic human actions captured from unconstrained environments. We described an action representation scheme using a rich set of descriptors computed from both local interest points and key point trajectories. We utilised an adaptive feature fusion method for combining trajectory based descriptor with spatio-temporal interest point based descriptors according to the detected camera movements. Crucially, we introduced a novel feature selection approach by formulating a discriminative MC-ΔLDA topic model for collaborative feature selection. The proposed framework outperforms most existing approaches on action recognition against realistic and unconstrained action recognition datasets.

References

[1] D. Andrzejewski, A. Mulhern, B. Liblit, and X. Zhu. Statistical debugging using latent topic models. In *ECML*, 2007.

[2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *ICCV*, volume 2, pages 1395–1402, 2005.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *CVPR*, 2009.

[5] N. P. Cuntoor and R. Chellappa. Trajectone, epitomic representation of human activities. In *CVPR*, 2007.

[6] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[7] C. Fanti, L. Zelnik-Manor, and P. Perona. Hybrid models for human motion recognition. In *CVPR*, 2005.

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

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

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

[11] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *ICCV*, 2009.

[12] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009.

[13] Thomas P. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft, 2000.

[14] O. Oshin, A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using randomised ferns. In *ICCV*, 2009.

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

[16] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, volume 3, pages 32–36, 2004.

[17] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, 2009.

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

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

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

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

[22] D. Zhang and G. Lu. A comparative study on shape retrieval using fourier descriptors with different shape signatures. In *JVCIR*, 2003.