Learning discriminative space-time actions from weakly labelled videos

BMVC 2011 Submission # ??

Abstract

Current *state-of-the-art* action classification methods derive action representations from the entire video clip in which the action unfolds, even though this representation may include parts of actions and scene context which are shared amongst multiple classes. For example, different actions involving the movement of the hands may be performed whilst walking, against a common background. In this work, we propose an action classification framework in which discriminative action *subvolumes* are learned in a weakly supervised setting, owing to the difficulty of manually labelling massive video datasets. The learned sub-action models are used to simultaneously *classify* video clips and to *localise* actions in space-time. Each subvolume is cast as a BoF instance in an MIL framework, which in turn is used to learn its class membership. We demonstrate quantitatively that the classification performance of our proposed algorithm is comparable and in some cases superior to the current *state-of-the-art* on the most challenging video datasets, whilst additionally estimating space-time localisation information.

1 Introduction

Human action recognition from video is becoming an increasingly prominent research area in computer vision, with far-reaching applications. On the web, the recognition of human actions will allow the organisation, search, description, and retrieval of information from the massive amounts of video data uploaded each day [14]. In every day life, human action recognition has the potential to provide a natural way to communicate with robots, and novel ways to interact with computer games and virtual environments.

In addition to the classical difficulties in computer vision of dealing with objects in the world with variations in illumination, viewpoint, background and part occlusions, human actions inherently possess a high degree of geometric and topological variability [3]. Furthermore, various human motions can carry the exact same meaning. For example, a jumping motion may vary in height, frequency and style, yet this is still the same action. It is therefore desirable for an action recognition system to generalise over actions in the same class, yet to discriminate between actions in different classes [21]. Despite the mentioned difficulties, significant progress has been made in learning and recognising human actions from videos [21, 28]. Whereas previously action recognition datasets included videos with single, staged human actions against homogeneous backgrounds [2, 22], more recently challenging uncontrolled movie data [17] and amateur video clips available on the Internet [14, 19] are being used to evaluate action recognition algorithms. These datasets contain human actions with

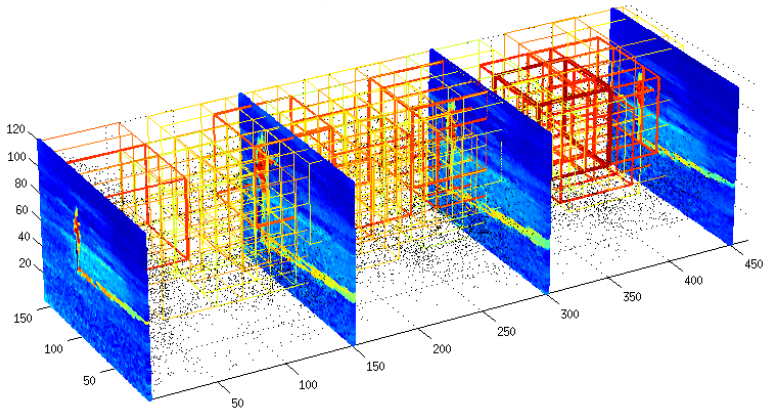


Figure 1: A boxing video sequence taken from the KTH dataset plotted in space-time. Discriminative action subvolumes learned in a max-margin multiple instance learning framework are plotted, with colour indicating their class membership strength. Notice that in this particular video, features were extracted from all space-time locations, and the camera was varying its zoom. In this case, since the context of the KTH dataset is not discriminative of the particular action, only subvolumes around the actor were selected as positive instances.

large variations in appearance, viewpoint, background clutter and camera motion, as they would appear in the real world.

Current *state-of-the-art* [7, 14, 18, 27] action clip classification methods derive an action representation from the entire video clip, even though this representation may contain motion and scene patterns pertaining to multiple action classes. For example, actions such as boxing and hand-clapping may be done whilst walking, standing or skipping, against a similar scene background. We therefore propose a framework in which action representations are derived from smaller portions of the video, subvolumes, which are used as learning primitives rather than the entire space-time video. In this way, more discriminative action parts may be selected which most characterise those particular types of actions. An example of learned action subvolumes is shown in Fig. 1.

2 Previous Work

The current *state-of-the-art* approach for the classification of challenging human action data has been the bag-of-features (BoF) on spatio-temporal volumes method [17, 26]. Typically in a first stage, local spatio-temporal structure and motion features are extracted from video clips and quantised to create a *visual vocabulary*. A query video clip is then represented using the frequency of occurring *visual words*, and classification is done using a χ^2-Kernel Support Vector Machine (SVM). The surprising success of the BoF method may be attributed to its ability to aggregate statistical information from local features, without regard to the detection of humans, body-parts or joint locations which are difficult to robustly detect in unconstrained action videos. However, its representational power diminishes with dataset difficulty (e.g. Hollywood2 dataset [20]) and an increased number of action classes (e.g. HMDB dataset [14]). This may be partly due to the fact that current BoF approaches repre-

sent entire video clips [27] or subsequences defined in a fixed grid [17]. Thus, many similar action parts, and background noise are also included in the histogram representation. By splitting up the video clip into overlapping subvolumes, a video clip is now represented as a bag of histograms, some of which are discriminative of the action and others which may hinder correct classification. A more robust action model can therefore be learned based on positive *subvolumes* of the space-time video. Moreover, the classification of subvolumes has the additional advantage of indicating *where* the action is happening in the space-time video.

In previous work, the BoF approach has been coupled with single frame person/action detection to gain more robust performance, and to estimate the action location [11, 16]. In contract, by learning discriminative action subvolumes from weakly-labelled videos, our proposed method will allow action localisation without using any training ground truth information, in a similar spirit to [2, 19]. Unlike previous work, however, we use a max-margin multiple instance learning (mi-SVM) framework to learn the latent class variables associated with each space-time action part.

Despite the availability of ground truth bounding box annotation, Viola *et al*. [25] applied a multiple instance learning framework to face detection. The improvement in recognition results as compared to a fully supervised framework suggests that there existed a more discriminative set of ground truth bounding boxes than those labelled by human observers. The difficulty in manual labelling arises from the inherent ambiguity in labelling objects or actions (bounding box scale, position) and the judgement, for each image/video, of whether the context is important for that particular example or not. A similar MIL approach was employed by Felzenszwalb and Huttenlocher [6], where possible object part bounding box locations were cast as latent variables. This allows the self-adjustment of the positive ground truth data, better aligning the learned objects filters during training. In action detection, Hu *et al*. used an MIL learning framework called SMILE-SVM [8], however this focused on detection of 2D action boxes, and required the approximate labelling of the frames and human heads in which the actions occur. In contrast, we propose to cast space-time subvolumes of cubic/cuboidial structure as latent variables, with the aim to capture salient action patches relevant to the human action.

In action clip classification it is assumed that only one action is occurring in each clip, and therefore the label of each action clip is known. This problem is inherently weakly-labelled since no approximate locations of the actions or ground truth action bounding boxes are available. Only the label of the whole video is known, and not the labels of individual parts of the action clip. Thus we propose to learn action subvolumes in a weakly-labelled, multiple instance learning (MIL) framework. Human actions are subsequently localised by detecting action instances in the query video, and mapping them to a final clip classification decision. Thereby, we propose an action recognition system that will be suitable for both clip classification and localisation in challenging video datasets, which does not necessitate labelling of action parts or locations.

The contributions of this work are as follows. We cast the conventionally supervised action clip classification pipeline into a weakly supervised setting, where action video clips are represented as bags of BoF instances with latent class variables. In order to learn the subvolume class labels, we bring multiple instance learning to 3D space-time videos. In this way, it is proposed that actions are better defined within a subvolume of a video clip rather than the whole video clip itself. This BoF-MIL approach will allow an action classifier to retrieve the location of the action as well as the label of the whole action clip, achieving *state-of-the-art* performance on the largest and most challenging video datasets.

3 Methods

The proposed action recognition system is composed of four main building blocks: i) the description of space-time video with histograms of Dense Trajectory features [27] (§3.1), ii) the representation of a video clip as a bag of subvolumes, each associated with a BoF histogram and a latent class label (§3.2), iii) the learning of positive subvolumes from weakly labelled training sequences with a max-margin multiple instance learning framework [1] (§ 3.3), and iv) mapping instance scores to bag scores with a standard SVM (§ 3.4).

3.1 Feature representation

A variety of interest point detectors (IPDs) have emerged for 3D spatio-temporal sequence representation [26]. Sparse features obtained using IPDs (Harris3D [15], Cuboid [5], Hessian [29]) are attractive because they allow compact video representations. With reference to human actions however, IPDs are not designed to capture smooth motions, and tend to also fire on highlights, shadows, object and video frame boundaries [10]. Furthermore, Wang *et al*. [26] demonstrated that dense sampling consistently outperformed IPDs in realistic video settings e.g. Hollywood2 dataset [20], suggesting that interest point detection for action recognition is still an open problem.

A plethora of video patch descriptors have been proposed for space-time volumes, mainly extended from their 2D counterparts: Cuboid [5], 3D-SIFT [23], HoG-HoF [17], HOG3D [12], extended SURF [29], C2-shape features [9], and Local Trinary Patterns [30]. More recently, Wang *et al*. [27] proposed Dense Trajectory features, which when combined with the standard BoF pipeline [26], outperformed the most recent *state-of-the-art* Learned Hierarchical Invariant features by Le *et al*. [18]. Therefore, even though this framework affords arbitrary features, we use the *Dense Trajectory* features of Wang *et al*. [27] to describe space-time video blocks.

Dense Trajectory features are extracted from a dense set of points, tracked at multiple spatial scales. A pruning stage eliminates static trajectories such as those found on homogeneous backgrounds, and spurious trajectories which may have drifted. The Dense Trajectory descriptor is formed by the sequence of displacement vectors in the trajectory, together with the HoG-HoF descriptor [17], and the motion boundary histogram (MBH) descriptor [4], computed over a local neighbourhood along the trajectory. The MBH descriptor represents the gradient of optical flow, which captures changes in the optical flow field, suppressing constant motions (e.g. camera panning), and capturing salient movements. Thus, Dense Trajectories capture the trajectory shape, appearance, and motion information [27]. Due to its success at action recognition in realistic settings, we use the BoF approach to describe video regions. The detailed BoF parameter settings are listed in Section 4.2.

3.2 Action clip representation

Unlike previous BoF action clip classification approaches which generate one histogram descriptor per action clip, either by counting the occurrences of visual words in the whole clip [22, 27], or by concatenating histograms from a spatial grid [17], we represent each video as a *bag* of possible histograms, as illustrated in Fig. 2. Each video volume is decomposed into multiple subvolumes, each of which is associated with a histogram of visual words and a latent variable representing its action class membership. This approach essentially converts

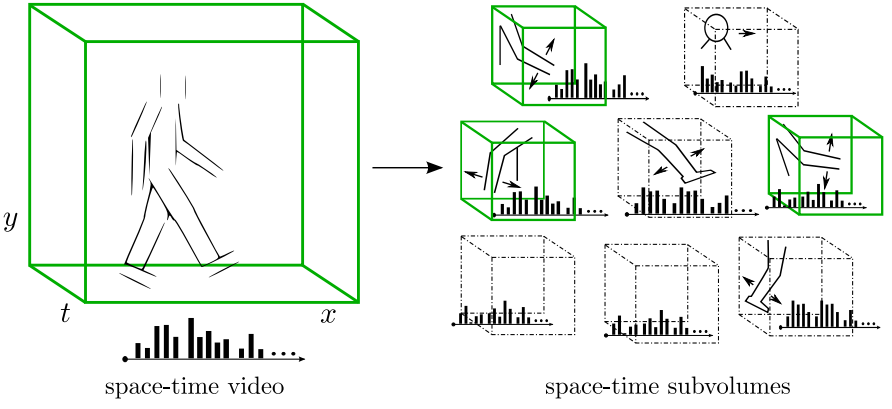


Figure 2: Instead of defining an action as a space-time pattern in an entire video clip (left), we propose to define an action by a collection of space-time action parts contained in video subvolumes (right). In the above illustration, one ground-truth action label is assigned to the entire space-time video, however, the labels of each action subvolume is initially unknown. Multiple instance learning is used to learn which subvolumes are particularly discriminative of the action (solid-line cubes), and which are not (dotted-line cubes).

the problem of whether an action clip contains a particular action, to whether smaller space-time fragments of the action clip contain the action. Thus each video is now represented by a bag of histograms, for which their individual class membership is initially unknown.

Consider training a classifier for a walking class action with all the action subvolumes generated from the walking videos in the training set. Initially all the instances/subvolumes are assumed to have the label of the parent bag/video, however whilst the negative labels always remain negative, the positive labels are allowed to switch to a negative label. Now consider a walking video instance whose feature distribution is also present in video cubes of other action classes. This kind of video instance will have a positive label if it originated from the walking videos, and a negative label from those similar instances drawn from the videos of the other classes (assuming a 1-vs-all classification approach). Thus, when these instances are reclassified in a future iteration, it is likely that their class label will switch to negative. As the class labels are updated in an iterative process, eventually only the most discriminative instances in each positive bag are retained as positive.

3.3 MIL-BoF action models

Recall that an action cube will be represented as a BoF histogram in a spatio-temporal volume. The task here is to learn an action model to represent each action class. In action classification datasets, an action class label is assigned to each video clip, assuming that one action occurs in each clip. This may be cast in a weakly labelled setting, where it is known that a positive example of the action exists within the clip, but the exact location of the action is unknown. If the label of the video clip/bag is positive, then it is assumed that a proportion of instances in the bag will also be positive. If the bag has a negative label, then all the instances in the bag must retain a negative label.

The learning task may be cast in a max-margin multiple instance learning framework

Algorithm 1 Heuristic algorithm proposed by [1] for solving mi-SVM.

Assign positive labels to instances in positive bags: $y_i = Y_I$ for $i \in I$
repeat
  Compute SVM solution $\mathbf{w}, b$ for instances in training set with estimated labels.
  Compute scores $f_i = \mathbf{w}^T x + b$ for all x_i in positive bags.
  Set $y_i = sgn(f_i)$ for all $i \in I$.
  for all positive bags B_I **do**
    if all instances are negative **then**
      find $i^* = \underset{i \in I}{\operatorname{argmax}} f_i$
      set $y_{i^*} = +1$
    end if
  end for
until class labels do not change
Output w, b

[1]. Let the training set $D = (\langle X_1, y_1 \rangle, ..., \langle X_n, y_n \rangle)$ consist of a set of bags $X_i = \{x_{i1}, ..., x_{in}\}$ and corresponding ground truth labels $y_i \in \{-1, +1\}$. Each example $x_{ij} \in \mathbb{R}$ represents the j^{th} BoF model in the bag, and whose label y_{ij} exists but is unknown for the positive training examples ($Y_i = +1$). The class label for each bag is positive if there exists at least one positive example in the bag, that is,

$$y_i = \max_j \{y_{ij}\}. \tag{1}$$

The max-margin mi-SVM learning problem results in a semi-convex optimisation problem, for which Andrews *et al*. proposed a heuristic approach [1], In mi-SVM, each example label is unobserved, and it maximises the usual soft-margin jointly over hidden variables and discriminant function:

$$\min_{y_i} \min_{\mathbf{w}, b, \xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i, \tag{2}$$

$$\text{subject to } \forall i: \quad y_i(\mathbf{w}^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad y_i \in \{-1, 1\}.$$

The heuristic algorithm proposed by Andrews *et al*. to solve the resulting mixed integer problem is laid out in Algorithm 1. Initially, the instance labels are set to match the bag labels, and a value for the latent variables is guessed with an estimate of the current model $\mathbf{w}$. Then, holding $\mathbf{w}$ fixed, the latent labels are re-estimated by setting $y_i = sgn(\mathbf{w}^T x_i + b)$. If all instances in a positive bag become negative, then the least negative instance in the bag is set to have a positive label, thus ensuring that there exists at least one positive example in a positive bag.

3.4 Making a classification decision

Thus far, the learned action models yield a classification score for each subvolume in a query video. Although this is desirable for action localisation, we focus here on action classification, and thus need to map the subvolume scores to a global video clip decision score. One solution is to take a threshold on some properties of the subvolumes, such as the number of positive subvolumes, the mean value of all subvolume scores, or a threshold on the

max/min decision scores from the clip. Since the number of subvolumes may vary greatly between videos, this is not trivially solved by normalisation. Consider a clearly performed action which only takes a small volume of the video clip. In an ideal case, there would be large scores for subvolumes containing the action, and low scores elsewhere. Clearly the normalised mean score/fraction of positive subvolumes would be very low, even though there was a valid action in the clip.

A simpler and more robust solution is to construct a linear classifier to separate decision score statistics obtained from positive and negative clips. Instead of learning a classifier from the decision scores directly, which will vary in number depending on the length of the clip, we consider six properties of the subvolumes in each clip: #pos, #neg, mean score, max score, min score, and learn a decision boundary in this constant dimensional space.

4 Experimental Evaluation & Discussion

In order to validate our action recognition system, we evaluated its performance on four challenging action datasets, namely the KTH, Hollywood2, YouTube, and HMDB datasets. All clips were down-sampled to a 160×120 resolution. In the following sections, a brief description of each dataset is laid out (§ 4.1), the baseline approach is detailed (§ 4.2, and we evaluate the MIL-BoF approach quantitatively (§ 4.3), and qualitatively (§ 4.4).

4.1 Datasets

The **KTH** dataset [22] contains 6 action classes (walking, jogging, running, boxing, waving, clapping) each performed several times by 25 actors, in four scenarios. We split the video samples into training and test sets according to [22], however we consider each video clip in the dataset to be a single action instance, and do not further slice the video into clean single action sequences.

The **YouTube** dataset [19] contains 11 action categories (basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking and walking with a dog), and presents several challenges due to camera motion, object appearance, scale, viewpoint and cluttered backgrounds. The 1600 video sequences are split into 25 groups, and we follow the author's evaluation procedure of 25-fold, leave-one-out cross validation.

The **Hollywood2** dataset [20] contains 12 action classes: answer phone, driving car, eating, fighting, getting out of car, hand-shaking, hugging, kissing, running, sitting down, sitting up, and standing up, collected from 69 different Hollywood movies. There are a total of 1707 action samples containing realistic, unconstrained human and camera motion. The dataset is divided into 823 training and 884 testing sequences, each from 5-25 seconds long.

The **HMDB** dataset [14] contains 51 action classes, with a total of 6849 video slips collected from movies, the Prelinger archive, YouTube and Google videos. Each action category contains a minimum of 101 clips. We use the same three train-test splits as the authors.

For each dataset, we present the current *state-of-the-art* result, our baseline BoF results, and our MIL-BoF on subvolumes variant. Moreover, we report the mean accuracy (mAcc), the average precision (mAP), and the mean F1-score (mF1) as evaluation metrics, revealing a more broad and realistic picture of the overall algorithm performance.

4.2 Baseline

The baseline approach uses the Dense Trajectory features [27] and the standard BoF pipeline described in [26], which achieved *state-of-the-art* results. We use the Dense Trajectory implementation provided by the author with trajectory length set to 15 frames, a neighbourhood size $N = 32$, and a spatio temporal grid size of $n_\sigma = 2, n_\tau = 3$, as per default [27]. In the BoF approach, a codebook was generated by randomly sampling 100,000 features and clustering them into 4000 visual words with k-means. Descriptors are assigned to their closest vocabulary word using the euclidean distance, and the resulting histograms of visual words are L1-normalised and used to represent each video clip. We report the performance achieved with both a standard linear-SVM and a χ^2-kernel SVM. Multi-class classification is done using the *one-vs-all* approach. We make no attempt to optimise the SVM regularisation parameters across the various datasets, and keep them constant at C=100 throughout, which has shown to give good performance experimentally.

4.3 MIL-BoF experimental setup

The same BoF parameters as the baseline are used for the MIL-BoF. The subvolumes are extracted from a regular grid with a grid spacing of 20 pixels in space and time . We collect results for a variation of cube [60-60-60], [80-80-80], [100-100-100] and cuboid [80-80-160], [80-160-80], [160-80-80] shaped subvolumes, where [x-y-t] denotes the dimensions of the part. We also allow for a certain type of cuboid to stretch along the total time duration of the clip, in a similar spirit to the weak geometrical, spatial pyramid approach of [17]. Initially, all the instances in each positive bag are set to have a positive label, as suggested by Andrews *et al*. [1].

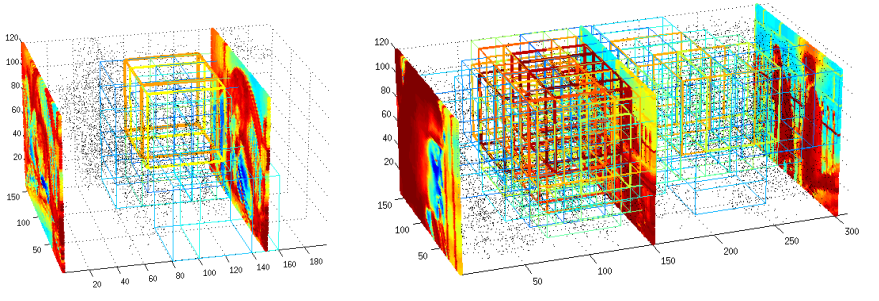
The decomposition of a video into multiple subvolumes, each with the same histogram dimensionality as used in the baseline, makes the learning problem at hand *large-scale*. Typical values for the number of instances generated from the KTH dataset range between 100,000-200,000. In practice calculating the full χ^2 kernel takes a prohibitively long time to compute. Recent work by Vedaldi and Zisserman on the homogeneous kernel map [24] demonstrates the feasibility of large scale learning with non-linear SVMs based on additive kernels, such as the χ^2 kernel. The map provides an approximate, finite dimensional feature representation in closed form, which gives a very good approximation of the desired kernel in a compact linear representation. The map parameters were set to N=2, and gamma=0.5, which gives a $2^n + 1$ dimensional approximated kernel map for the χ^2 kernel. Similarly to the baseline, we keep the SVM parameters constant across all datasets at C=10 for the linear SVM and C=0.1 for the χ^2 kernel SVM, which have proven to give good results in practice. We report the MIL-BoF results obtained using i) linear-SVM to learn the latent class labels and action model (LL-SVM), ii) linear-SVM to learn latent labels and χ^2 kernel SVM for the action model (LK-SVM), and iii) χ^2 kernel approach for both steps (KK-SVM). The quantitative results are shown in Table 1.

4.4 Discussion

On the KTH dataset the baseline surpassed the *state-of-the-art* despite the dataset not being cleanly split into action slices. This may primarily be due to the Dense Trajectory features which prune out unwanted portions of the videos where no motion occurs. However, the best result was achieved by the MIL variant with 96.3% accuracy. On the YOUTUBE dataset,

Table 1: Comparison of our approach to the baseline and soda approaches

Dataset	KTH			YOUTUBE			HOHA2			HMDB		
Perf. measure	mAcc	mAP	mF1	mAcc	mAP	mF1	mAcc	mAP	mF1	mAcc	mAP	mF1
State-of-the-art	94.53[1,2]		-	**84.2**[2]	-	-	-	**58.3**[2]	-	-	23.18[1]	-
L-BoF	93.98	95.38	89.08	64.14	76.75	58.26	37.08	39.05	**39.17**	22.31	19.02	19.98
K-BoF	95.37	96.48	93.99	76.03	79.33	57.54	**39.04**	**48.73**	32.04	**31.53**	**31.39**	21.36
LL-MIL-BoF 60	94.44	96.59	93.96	72.38	79.25	63.99	0.00	0.00	0.00	21.74	21.14	19.38
KK-MIL-BoF 60	94.91	96.48	94.22	73.40	81.04	70.04	0.00	0.00	0.00	27.64	26.26	23.08
LK-MIL-BoF 60	**96.30**	97.01	94.06	73.34	81.57	69.41	0.00	0.00	0.00	26.70	26.72	23.47
LL-MIL-BoF 80	94.44	96.61	**95.09**	74.17	79.61	66.08	35.69	40.03	38.60	22.37	21.72	19.59
KK-MIL-BoF 80	95.37	**97.02**	94.84	77.54	83.86	73.94	37.28	44.18	37.45	28.69	29.03	25.28
LK-MIL-BoF 80	94.91	96.84	94.63	76.41	83.22	72.38	34.55	42.75	38.49	28.60	29.40	**25.75**
LL-MIL-BoF 100	94.44	96.63	93.32	75.31	80.27	68.49	32.53	37.10	32.44	21.04	19.93	17.71
KK-MIL-BoF 100	93.52	96.53	93.65	78.60	85.32	76.29	37.43	40.72	32.31	27.51	28.62	23.93
LK-MIL-BoF 100	93.52	96.35	93.91	**78.86**	**85.40**	**76.28**	34.84	42.03	36.90	27.32	28.76	24.01



(a) actioncliptest00135 (b) actioncliptest00058

Figure 3: Action localisation results on two challenging videos from the Hollywood2 dataset, which we encourage the reader to watch in support of this figure. The colour of the boxes indicates the positive rank score of the subvolume belonging to a particular class. (a) In actioncliptest00135, the phoning action does not occur until half way through the video, correctly identified by the position in space-time of the detected subvolumes. (b) In actioncliptest00058, a woman is getting out of the car, however this action occurs in the middle of the video and not at the beginning or end, as indicated by the detected subvolumes.

despite the baseline not being able to reach the *state-of-the-art* we report a 5%, 8%, and 18% increase in accuracy, average precision and F1 score respectively when compared to the baseline, demonstrating the ability to learn mode robust action models. On the HMDB dataset, we report baseline performance superior to the current *state-of-the-art* however the MIL-BoF outperforms the baseline by 4% on the F1 score, which weights precision and recall equally. Hollywood2 data is still being computed. padding text padding text padding text padding text padding text padding text padding text padding text padding text padding text padding text padding text. One reason why the MIL does not always outperform the baseline may be because it is not guaranteed to converge to the optimal solution. Despite this, there is still an added advantage of being able to estimate the action's space-time location as shown in Fig 4.4. Since in our approach the detection primitives are space-time subvolumes, there is no need to perform spatial and temporal detection separately [19]. Each subvolume is associated with a location in the video, and a decision score for each action class.

5 Conclusion

We proposed a novel MIL-BoF approach to action clip classification and localisation based on the detection of space-time subvolumes. By learning the latent class variables with multiple instance learning, more robust action models may be constructed, and which may be used for both action clip classification and localisation. The experimental results demonstrate the the our method is comparable and exceeds the *state-of-the-art* on some challenging datasets. In future, we will focus on quantifying the action localisation performance, and the use of mixed subvolume models.

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pages 561–568, 2003.

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

[3] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Topology-invariant similarity of nonrigid shapes. *Int. Journal of Computer Vision*, 81(3):281–301, 2009.

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. European Conf. Computer Vision*, 2006.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[7] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Proc. Int. Conf. Computer Vision*, pages 925–931, 2009.

[8] Yuxiao Hu, Liangliang Cao, Fengjun Lvand Shuicheng Yan, Yihong Gong, and Thomas S. Huang. Action detection in complex scenes with spatial and temporal ambiguities. In *Proc. Int. Conf. Computer Vision*, pages 128–135, 2009.

[9] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *Proc. Int. Conf. Computer Vision*, pages 1–8, 2007.

[10] Y. Ke, R. Sukthandar, and M. Hebert. Volumetric features for video event detection. *Int. Journal of Computer Vision*, 88(3):339–362, 2010.

[11] A. Kläser, M. Marszałek, C. Schmid, and A. Zisserman. Human focused action localization in video. In *International Workshop on Sign, Gesture, Activity*, 2010.

[12] Alexander Klaser, Marcin Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3D-gradients. In *Proc. British Machine Vision Conference*, 2008.

[13] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 2046–2053, 2010.

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *Proc. Int. Conf. Computer Vision*, 2011.

[15] I. Laptev and T Lindeberg. Space-time interest points. In *Proc. Int. Conf. Computer Vision*, 2003.

[16] Ivan Laptev and Patrick Pérez. Retrieving actions in movies. In *Int. Conf. on Computer Vision*, pages 1–8, 2007.

[17] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[18] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recignition with independent subspace analysis. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 3361–3368, 2011.

[19] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognising realistic actions from videos "in the wild". In *Proc. British Machine Vision Conference*, pages 1996–2003, 2009.

[20] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 2929–2936, 2009.

[21] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28:976–990, 2010.

[22] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *IEEE Int. Conf. on Pattern Recognition*, pages 32–36, 2004.

[23] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In *Proc. ACM Multimedia*, 2007.

[24] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 3539–3546, 2010.

[25] P. Viola, J. Platt, and C. Zhang. Muliple instance boosting for object detection. In *Advances in Neural Information Processing Systems*, 2005.

[26] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. British Machine Vision Conference*, 2009.

[27] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action Recognition by Dense Trajectories. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3169–3176, 2011.

[28] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224 – 241, 2011.

[29] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proc. European Conf. Computer Vision*, pages 650–663, 2008.

[30] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *Proc. Int. Conf. Computer Vision*, 2009.