Learning discriminative space-time actions from weakly labelled videos

Michael Sapienza
michael.sapienza-2011@brookes.ac.uk
Fabio Cuzzolin
fabio.cuzzolin@brookes.ac.uk
Philip H.S. Torr
philiptorr@brookes.ac.uk

Brookes Vision Group
Oxford Brookes University
Oxford, UK
cms.brookes.ac.uk/research/visiongroup

Abstract

Current *state-of-the-art* action classification methods extract feature representations from the entire video clip in which the action unfolds, however this representation may include irrelevant scene context and movements which are shared amongst multiple action classes. For example, a waving action may be performed whilst walking, however if the walking movement and scene context appear in other action classes, *then they should not be included* in a waving movement classifier. In this work, we propose an action classification framework in which more discriminative action *subvolumes* are learned in a weakly supervised setting, owing to the difficulty of manually labelling massive video datasets. The learned models are used to simultaneously *classify* video clips and to *localise* actions to a given space-time subvolume. Each subvolume is cast as a bag-of-features (BoF) instance in a multiple-instance-learning framework, which in turn is used to learn its class membership. We demonstrate quantitatively that even with single fixed-sized subvolumes, the classification performance of our proposed algorithm is superior to the *state-of-the-art* BoF baseline on the majority of performance measures, and shows promise for space-time action localisation on the most challenging video datasets.

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 being subject to the usual nuisance factors such as variations in illumination, viewpoint, background and part occlusions, human actions inherently possess a high degree of geometric and topological variability [3]. Various human motions can carry the exact same meaning. For example, a jumping motion may vary in height, frequency and style, yet still be the same action. Action recognition systems need then to generalise over actions in the same class, while discriminating between actions of different classes [21].

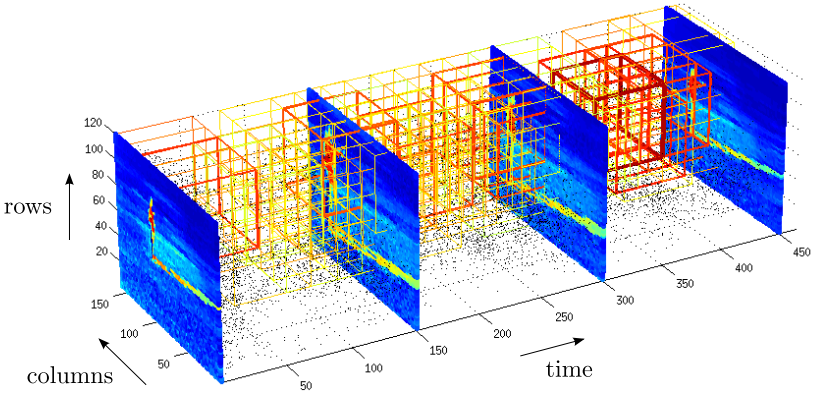


Figure 1: *A boxing video sequence taken from the KTH dataset [22] plotted in space and time. Notice that in this particular video, the camera zoom is varying with time, and features (black dots) were extracted from all space-time locations. Overlaid on the video are discriminative cubic action subvolumes learned in a max-margin multiple instance learning framework (§ 2.2), with colour indicating their class membership strength. Since the scene context of the KTH dataset is not discriminative of the particular action, only subvolumes around the actor were selected as positive instances (best viewed in colour).*

Despite these difficulties, significant progress has been made in learning and recognising human actions from videos [21, 28]. Whereas early 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 [12, 19] are being used to evaluate action recognition algorithms.

Current *state-of-the-art* [7, 14, 18, 27] action clip classification methods derive an action's representation from an 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 performed whilst walking, standing or skipping, against a similar scene background. The presence of similar sub-motions or scene parts can therefore negatively affect recognition rates. We therefore propose a framework in which action models are derived from smaller portions of the video volume, 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.

Previous Work

The current *state-of-the-art* algorithms for the classification of challenging human action data are based on the bag-of-features (BoF) on spatio-temporal volumes approach [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 the 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 for 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

represent 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 instead represented as a bag of histograms, some of which are discriminative of the action at hand ('positive' subvolumes) while others ('negative' ones) may hinder correct classification. A more robust action model can therefore be learned based on these *positive subvolumes* in the space-time video. Moreover, the classification of subvolumes has the additional advantage of indicating *where* the action is happening within the 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 [12, 16]. In contrast, by learning discriminative action subvolumes from weakly-labelled videos, the method we propose *allows action localisation without using any training ground truth information*, in a similar spirit to [2, 19]. Unlike previous work, however, we select discriminative feature histograms and not the explicit features themselves. Moreover, instead of using a generative approach such as pLSA [30], we use a max-margin multiple instance learning (mi-SVM) framework to handle the latent class variables associated with each space-time action part.

Some insight to MIL comes from its use in the context of face detection [25]. Despite the availability of ground truth bounding box annotation, the improvement in detection results when compared to those of a fully supervised framework suggested 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] for object detection in which possible object part bounding box locations were cast as latent variables. This allowed the self-adjustment of the positive ground truth data, better aligning the learned object filters during training. In action detection, Hu *et al*. used an MIL learning framework called SMILE-SVM [8]; however, this focused on the 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 casting the 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 only the label of each action clip is known, and not the labels of individual parts of the action clip. Thus, this problem is *inherently weakly-labelled*, since no approximate locations of the actions or ground truth action bounding boxes are available. That is why we propose to learn action subvolumes in a weakly-labelled, multiple instance learning (MIL) framework. Human action classification is then achieved by the recognition of action instances in the query video, after devising a sensible mapping from recognition scores to the final clip classification decision.

Contributions

The contributions of this work are as follows: i) We cast the conventionally supervised BoF action clip classification approach into a weakly supervised setting, where clips are represented as bags of histogram instances with latent class variables. ii) In order to learn the subvolume class labels, we apply multiple instance learning to 3D space-time videos (§2.2), as we maintain that actions are better defined within a subvolume of a video clip rather than the whole video clip itself. iii) Finally we propose a mapping from *instance* decisions learned in the mi-SVM approach to *bag* decisions (§2.3), as a more robust alternative to the

current bag margin MIL approach of taking the sign of the maximum margin in each bag. This allows our MIL-BoF approach to learn the labels of each individual subvolume in an action clip, *as well as the label of the action clip as a whole*. The resulting action recognition system is suitable for both clip classification and localisation in challenging video datasets, without requiring the labelling of action part locations.

2 Methodology

The proposed action recognition system is composed of three main building blocks: i) the description of space-time videos via histograms of Dense Trajectory features [27] (§2.1), ii) the representation of a video clip as a "bag of subvolumes", and the learning of positive subvolumes from weakly labelled training sequences within a max-margin multiple instance learning framework (§ 2.2), and iii) the mapping of instance/subvolume scores to bag/clip scores by learning a hyperplane on instance margin features (§ 2.3).

2.1 Feature representation

A variety of interest point detectors (IPDs) are being used for 3D spatio-temporal sequence representation [26]. Whereas sparse features obtained using IPDs (Harris3D [15], Cuboid [5], Hessian [29]) allow compact video representations, IPDs are not designed to capture smooth motions associated with human actions, and tend to fire on highlights, shadows, and video frame boundaries [7, 10]. Furthermore, Wang *et al*. [26] demonstrated that dense sampling outperformed IPDs in real video settings such as the Hollywood2 dataset [20], implying 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 derived from their 2D counterparts: Cuboid [5], 3D-SIFT [23], HoG-HoF [17], HOG3D [11], extended SURF [29], and C2-shape features [9]. More recently, Wang *et al*. [27] proposed Dense Trajectory features, which when combined with the standard BoF pipeline [26], outperformed the recent Learned Hierarchical Invariant features by Le *et al*. [18]. Therefore, even though this framework is independent from the choice of features, we use the *Dense Trajectory* features of Wang *et al*. [27] to describe space-time video blocks.

The Dense Trajectory features are extracted densely from a video 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 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 the optical flow, and captures changes in the optical flow field, suppressing constant motions (e.g. camera panning), and capturing salient movements. Thus, Dense Trajectories capture a trajectory's shape, appearance, and motion information [27]. Due to its success with action recognition in realistic settings, we use the BoF approach to describe space-time subvolumes. The detailed BoF parameter settings are listed in Section 3.

2.2 MIL-BoF action models

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

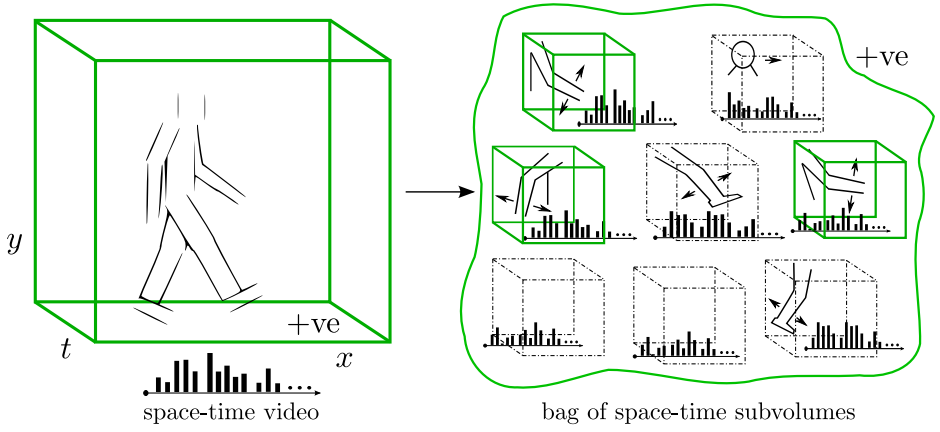


Figure 2: *Instead of defining an action as a space-time pattern in an entire video clip (left), we propose to define an action as a collection of space-time action parts contained in video subvolumes (right). One ground-truth action label is assigned to the entire space-time video, while the labels of each action subvolume are 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).*

[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 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.

The task here is to learn the class membership of each subvolume and 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, of which the pattern/instance margin formulation [1] is best suited for space-time subvolume localisation. Let the training set $D = (\langle X_1, Y_1\rangle, \ldots, \langle X_n, Y_n\rangle)$ consist of a set of bags $X_i = \{x_{i1}, \ldots, x_{im_i}\}$ of different length and corresponding ground truth labels $Y_i \in \{-1, +1\}$. Each instance $x_{ij} \in \mathbb{R}$ represents the j^{th} BoF model in the bag: its label y_{ij} does exist, but is unknown for the positive bags ($Y_i = +1$). The class label for each bag is positive if there exists at least one positive example in the bag, that is, $Y_i = \max_j\{y_{ij}\}$. Therefore the task of the mi-MIL is to recover the latent variable y_{ij} of every instance in the positive bags, and to simultaneously learn an SVM instance model $\langle \mathbf{w}, b\rangle$ to represent each action class.

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 we maximise the usual soft-margin jointly over hidden variables

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

STEP 1. Assign positive labels to instances in positive bags: $y_{ij} = Y_i$ for $j \in i$
repeat
  STEP 2. Compute SVM solution $\langle \mathbf{w}, b \rangle$ for instances with estimated labels y_{ij}.
  STEP 3. Compute scores $f_{ij} = \mathbf{w}^T x_{ij} + b$ for all x_{ij} in positive bags.
  STEP 4. Set $y_{ij} = sgn(f_{ij})$ for all $j \in i$, $Y_i = 1$.
  for all positive bags X_i **do**
    if $\sum_{j \in i}(1 + y_{ij})/2 == 0$ **then**
      STEP 5. Find $j^* = \underset{j \in i}{\operatorname{argmax}} f_{ij}$, and set $y^*_{ij} = +1$
    end if
  end for
until class labels do not change
Output $\mathbf{w}, b$

and discriminant function:

$$\min_{y_{ij}} \min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{ij} \xi_{ij}, \tag{1}$$

$$\text{subject to } \forall i, j: \quad y_{ij}(\mathbf{w}^T x_{ij} + b) \geq 1 - \xi_{ij}, \quad \xi_{ij} \geq 0, \quad y_{ij} \in \{-1, 1\},$$

where $\mathbf{w}$ is the normal to the separating hyperplane, b is the offset, and ξ_{ij} are slack variables for each instance x_{ij}.

The heuristic algorithm proposed by Andrews *et al*. to solve the resulting mixed integer problem is laid out in Algorithm 1. 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 (STEP 1). Next, a walking action model estimate $\langle \mathbf{w}, b \rangle$ is found using the imputed labels y_{ij} (STEP 2), and scores for each subvolume in the bag are estimated with the current model (STEP 3). Whilst the negative labels always remain negative, the positive labels may retain their current label, or switch to a negative label (STEP 4). If, however, all instances in a positive bag become negative, then the least negative instance in the bag is set to have a positive label (STEP 5), thus ensuring that there exists at least one positive example in each positive bag.

Now consider a walking video instance whose feature distribution is also present in video subvolumes 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*.

2.3 A mapping from instance to bag labels in the MIL framework

The instance margin formulation detailed in the previous section aims at recovering the latent variables of all instances in each positive bag. When recovering the optimal labelling and the optimal hyperplane, *all* the positive and negative instances in a positive bag are considered. Thus, a query subvolume is predicted to have a label $\hat{y}_{ij} = \text{sgn}(\mathbf{w}^T x_{ij} + b)$. A similar MIL

approach called the "bag margin" formulation is typically adopted to predict the label of the bag from its instances. This approach, unlike the instance margin formulation, only considers the "most positive" pattern in each positive bag. Therefore, predictions take the form:

$$\hat{Y}_i = \text{sgn} \max_{j \in i}(\mathbf{w}^T x_{ij} + b). \tag{2}$$

In order to avoid retrieving the bag label by performing a bag-margin iterative procedure similar to the one detailed in section 2.2, we propose a simple and robust alternative method to predict the bag margins from the instance margins. One solution is use the same max decision rule in (2), or to take a threshold on some quantities, such as the number of positive subvolumes, or the mean value of all subvolume scores. Since the number of subvolumes may vary greatly between videos, this cannot be trivially solved by normalisation. Consider a neatly 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 hyperplane separating instance margin features $\mathcal{F}_i$ obtained from the positive and negative bags. Instead of learning a classifier from the margin values directly, which *will vary in number greatly* depending on the number of instances in each clip, we consider six features of the instance/subvolume margins $f_{ij} = \mathbf{w}^T x_{ij} + b$ in each clip i:

$$\mathcal{F}_i = \left[\#pos,\ \#neg,\ \#pos/\#neg,\ 1/n\textstyle\sum_j(f_{ij}),\ \max_{j \in i}(f_{ij}), \min_{j \in i}(f_{ij})\right], \tag{3}$$

where $\#pos$ and $\#neg$ are the number of positive/negative instances in each bag respectively. Thus, the instance margins are mapped to a six-dimensional feature vector, and a decision boundary is learned in this *constant dimensional space*.

3 Experimental Evaluation & Discussion

In order to validate our action recognition system, we evaluated its performance on four challenging action datasets, namely the KTH, YouTube, Hollywood2 and HMDB datasets. We give a brief overview of the datasets and the baseline pipeline, followed by the details of our MIL-BoF experimental setup (§ 3.1), and an ensuing discussion (§ 3.2).

Datasets and experimental setup

The *KTH* dataset [22] contains 6 action classes each performed by 25 actors, in four scenarios. *We split the video samples into training and test sets as in* [22]; however, we consider each video clip in the dataset to be a single action sequence, and do not further slice the video into clean, smaller action clips. This shows the robustness of our method to longer video sequences which include noisy segments in which the actor is not present.

The *YouTube* dataset [19] contains 11 action categories 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 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, as in* [20], each from 5-25 seconds long.

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

In order to make the comparison across different datasets fairer, all clips were downsampled to a common 160×120 resolution. For each dataset, we present both the state-of-the-art result as reported in the literature, and the baseline BoF results in our own implementation, to which we compare our MIL-BoF on subvolumes framework. As performance measures, we report the accuracy (Acc) calculated as the #correctly classified testing clips/#total testing clips, the average precision (AP) which considers the ordering in which the results are presented, and the F1-score which weights recall and precision equally and is calculated as: $F1 = (2 \times \text{recall} \times \text{precision})/(\text{recall} + \text{precision})$. Unlike previous work, we are *the first to report all three performance measures for each dataset*, to give a more complete picture of the overall algorithm performance.

Baseline BoF algorithm
We have implemented the baseline BoF approach described in [26] to ensure a fair comparison between BoF and MIL-BoF. A codebook is generated by randomly sampling 100,000 features and clustering them into 4000 visual words by k-means. Descriptors are assigned to their closest vocabulary word using the Euclidean distance, and the resulting histograms of visual words used to represent each video clip. We report the performance achieved using a χ^2 kernel SVM [26], and perform multi-class classification using the *one-vs-all* approach. We fix the histogram normalisation to the $L1$-norm, and *we make no attempt to optimise the SVM regularisation parameters across the various datasets*. We keep C=100 throughout, the same value used by [18, 26].

3.1 MIL-BoF experimental setup

The same BoF setup as the baseline has been used for the MIL-BoF approach. Subvolumes are extracted from a regular grid with a grid spacing of 20 pixels in space and time. Results are reported for a number of different MIL-BoF models, each characterised by different cube-[60-60-60], [80-80-80], [100-100-100] or cuboid-[80-80-160], [80-160-80], [160-80-80] shaped subvolumes, where [x-y-t] denotes the dimensions of the subvolume. In addition, we also allow for a certain type of cuboid to stretch along the total time duration of the clip [80-80-end], in a similar spirit to the weak geometrical, spatial pyramid approach of [17].

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=0.1, which has proven to give good results in practice. The quantitative results are shown in Table 1.

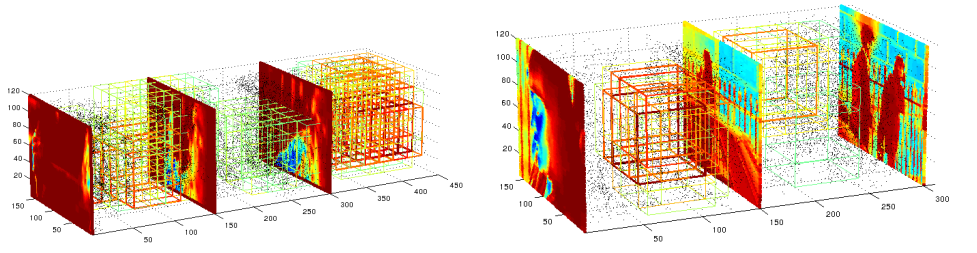
Table 1: Quantitative results from BoF and MIL-BoF methods.

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[8, 9]		–	**84.2**[9]	–	–	–	**58.3**[9]	–	23.18[9]	–	–
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
MIL-BoF 60-60-60	94.91	96.48	94.22	73.40	81.04	70.04	38.49	43.49	**39.42**	27.64	26.26	23.08
MIL-BoF 80-80-80	95.37	**97.02**	94.84	77.54	83.86	73.94	37.28	44.18	37.45	28.69	29.03	25.28
MIL-BoF 100-100-100	93.52	96.53	93.65	78.60	85.32	76.29	37.43	40.72	32.31	27.51	28.62	23.93
MIL-BoF 80-80-160	**96.76**	96.74	95.78	**80.39**	86.06	**77.35**	37.49	41.97	33.66	28.17	29.55	**25.41**
MIL-BoF 160-80-80	96.30	96.58	94.44	79.05	85.03	76.07	36.92	42.08	32.11	28.98	30.50	24.76
MIL-BoF 80-160-80	95.83	96.62	94.41	78.31	84.94	75.74	37.84	42.61	35.33	28.71	28.82	25.26
MIL-BoF 80-80-end	**96.76**	96.92	**96.04**	79.27	**86.10**	75.94	**39.63**	43.93	35.96	29.67	30.30	25.22

3.2 Discussion

On the **KTH** dataset the *MIL-BoF approach surpassed the baseline BoF in all three performance measures*, demonstrating a clear advantage of representing videos with subvolumes on this dataset. Common scene and motion elements were pruned by the multiple-instance learning as shown in Fig 1, resulting in a stronger action classifier per class. Contrary to our expectations, both the BoF and MIL-BoF surpassed the *state-of-the-art* accuracy, which may be attributed to using the whole action videos rather than clean action slices during training. The best result was achieved using a subvolume model more extended in time than in space [80-80-160], that achieved 96.76% accuracy. Similarly on the **YOUTUBE** dataset, the MIL-BoF framework outperformed the baseline BoF on *all* performance measures, achieving a 4.36%, 6.73%, and 19.81% *increase* in accuracy, average precision and F1 score respectively. This demonstrates the MIL-BoF ability to learn more robust action models on challenging YouTube data. The MIL-BoF approach did not improve the AP compared to the baseline on the **HOHA2** dataset, however, this was made up for by a 0.59% increase in Accuracy and a 7.38% improvement on the F1 score, *which weights precision and recall equally*. On the **HMDB** dataset, we report a BoF baseline performance superior to the current *state-of-the-art*. Similarly to the Hollywood2 dataset, our MIL-BoF approach outperforms the BoF baseline on the F1 score, in this case by 4.05%. In accord with observations in [17], we achieve good results with subvolume primitives in which there is no temporal subdivision of the sequence [80-80-end], however, we show that a temporal subdivision of the action sequence [80-80-160] can in fact result in a sizable improvement over considering no temporal subdivision at all, as may be seen in the F1-scores of the YOUTUBE [80-80-160] and HOHA2 [60-60-60] dataset.

Our MIL-BoF algorithm is not guaranteed to converge to the optimal solution, and may be one reason why it did not improve over the baseline Accuracy and AP on the HMDB dataset. However, bear in mind that the full χ^2 kernel is calculated for the BoF baseline whilst the linear approximation [24] was used in the MIL-BoF. We expect the results to improve further in the case of full resolution videos. Moreover, due to the large computational cost associated with space-time subvolumes, the full potential of our algorithm has yet to be realised, when a *more general mixture of subvolume primitives is tailored automatically for each action class*. Despite these current setbacks, the MIL-BoF method still outperforms the baseline BoF method in all performance measures on the KTH and YOUTUBE dataset, whilst outperforming the HOHA2 and HMDB on the F1 score, even with fixed-sized subvolumes. Finally, in addition to clip classification, the MIL-BoF method is able to localise challenging actions in space-time, such as the DriveCar and GetOutOfCar actions in the HOHA2 dataset shown in Fig 3(a) & 3(b).



(a) actioncliptest00032 (b) actioncliptest00058

Figure 3: *Action localisation results on two challenging videos from the Hollywood2 dataset, which we encourage the reader to watch in addition to this figure. The colour of each box indicates the positive rank score of the subvolume belonging to a particular class (green-red). (a) Actioncliptest00032 begins with two people chatting in a car. Half-way in, the camera shot changes to a view from the roof of the car. Finally the shot returns to the two people, this time the steering wheel is visible and the driving action is evident. This is reflected by the densely detected, high scoring subvolumes towards the end of actioncliptest00032. (b) In actioncliptest00058, a woman is getting out of her car, however, this action occurs in the middle of the video and not at the beginning or end, as indicated by the detected subvolumes.*

4 Conclusion

We proposed a novel MIL-BoF approach to action clip classification *and* localisation based on the recognition of space-time subvolumes. By learning the subvolume latent class variables with multiple instance learning, more robust action models may be constructed and used for action localisation in space and time or action clip classification via our proposed mapping from instance to bag decision scores. The experimental results demonstrate that the MIL-BoF method achieves comparable performance or improves on the BoF baseline on the most challenging datasets. In the future, we will focus on generalising the MIL-BoF approach by learning a *mixture of subvolume primitives* tailored for each action class, and incorporating geometric structure by means of *pictorial star models*.

Acknowledgements: This work was partially funded under EPSRC research grant EP/I018719/1, and by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.
P. H.S. Torr is in receipt of Royal Society Wolfson Research Merit Award.

References

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

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. 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*, pages 428–441, 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] Y. Hu, L. Cao, F.L.S. Yan, Y. Gong, and T.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, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *Proc. British Machine Vision Conference*, pages 99.1–99.10, 2008.

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

[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*, pages 2556–2563, 2011.

[15] I. Laptev and T Lindeberg. Space-time interest points. In *Proc. Int. Conf. Computer Vision*, pages 432–439, 2003.

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

[17] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[18] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.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] J. Liu, J. Luo, and M. Shah. Recognising realistic actions from videos "in the wild". In *Proc. British Machine Vision Conference*, pages 1996–2003, 2009.

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

[21] R. 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*, pages 357–360, 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*, pages 1417–1426, 2005.

[26] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. British Machine Vision Conference*, pages 124.1–124.11, 2009.

[27] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action Recognition by Dense Trajectories. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3169–3176, 2011.

[28] D. Weinland, R. Ronfard, and E. 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] S. Wong, T. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1–6, 2007.