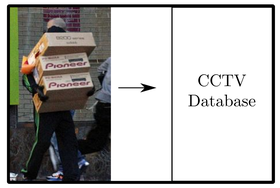
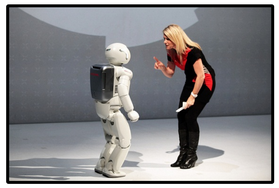
# Learning discriminative space-time actions from weakly labelled videos

Michael Sapienza, Fabio Cuzzolin and Philip H.S. Torr
Oxford Brookes University, Oxford, UK

## Motivation




**Billions of videos require:**
- Organization, Search
- Description, Retrieval



**Recognising actions for:**
- Human Robot Interaction
- Gaming, Virtual Reality

## State-of-the-art

- The space-time **bag-of-features (BoF)** approach is the most popular pipeline for challenging human action data [5, 9], however classification performance diminishes with dataset difficulty (e.g. HMDB [4]).
- State-of-the-art methods [8, 2, 4, 6] derive action representations from an entire video clip, **even though this may contain motion and scene patterns pertaining to multiple action classes**.
- Different actions that have similar motions may lead to confusion between classes.

## Our approach

- Human actions may naturally be described as a collection of parts.

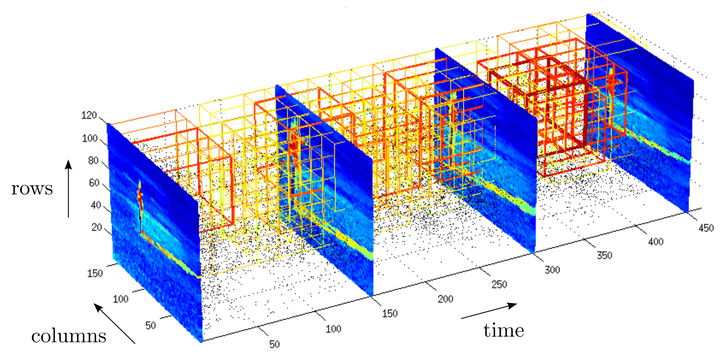



Figure: *A training video sequence taken from the KTH dataset [7] plotted in space and time. Overlaid on the video are discriminative cubic action subvolumes learned in a max-margin multiple instance learning framework, with colour indicating their class membership strength.*

- In our framework, 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**.

## Method

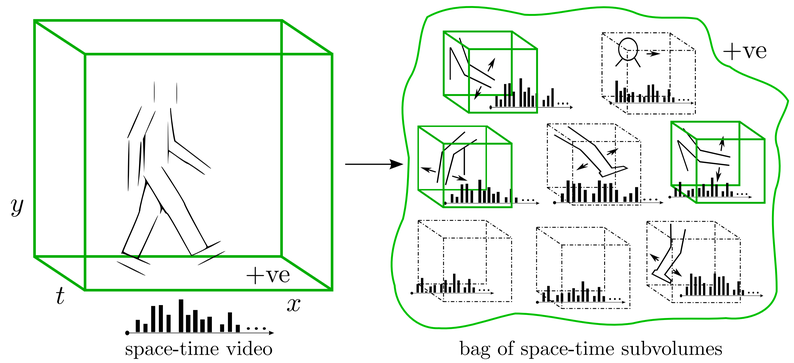
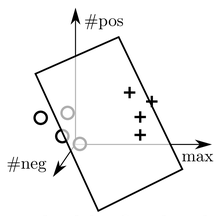



Figure: *Instead of defining an action as a space-time pattern in an entire video clip (left), we propose to define an action as a collection of space-time action parts contained in video subvolumes (right). The labels of each action subvolume are initially unknown. Multiple instance learning is used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

- i) Cast conventionally supervised BoF action clip classification into a weakly supervised setting.
- ii) Video clips are represented as bags of histogram instances with latent class variables.
- iii) Apply multiple instance learning (MIL) to learn the subvolume class labels.
- iv) Map *instance* decisions learned in the mi-SVM approach to *bag* decisions by learning a hyperplane separating instance margin features $\mathcal{F}_i$ in pos/neg bags.




- **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.**

## MIL-BoF

- The task of the mi-MIL is to recover the latent class 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.
- In mi-SVM, each example label is unobserved, and we maximise the usual soft-margin jointly over hidden variables and discriminant function [1]:

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

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 $\xi_{ij}$ are slack variables for each instance $x_{ij}$.

## Experimental Setup

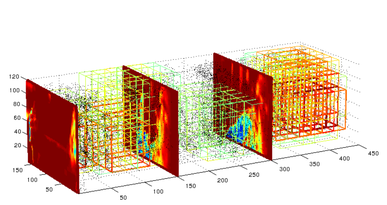
- 4 Challenging action datasets: **KTH** (6 classes), **YouTube** (11 classes), **Hollywood2** (12 classes), **HMDB** (51 classes).
- **Baseline**: Standard BoF pipeline [9].
- **Our approach**: MIL-BOF models characterised by various cube-[60-60-60], [80-80-80], [100-100-100] or cuboid-[80-80-160], [80-160-80], [160-80-80] shaped subvolumes [x-y-t].

## Results

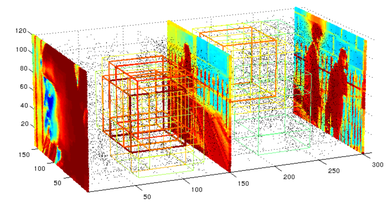
Table: Quantitative action clip **classification** results from BoF and MIL-BoF methods.

| **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[3, 2] | – | | **84.2**[8] | – | – | – | **58.3**[8] | | 23.18[4] | | – |
| 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 |

Figure: Qualitative action clip **localization** results on challenging Hollywood2 dataset.



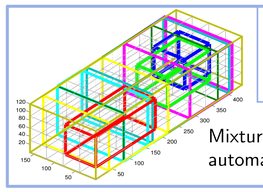
(a) Action: DriveCar - test video



(b) Action: GetOutOfCar - test video

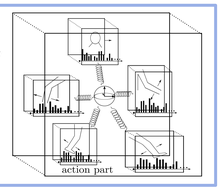
## Conclusion

- Even with fixed-sized subvolumes, MIL-BoF achieves comparable & superior performance to BoF baseline on most performance measures.



**Future work**:
Incorporate geometric structure by means of pictorial star models

Mixture of subvolume primitives tailored automatically for each action class




## References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2003.
[2] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*, 2009.
[3] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.
[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *ICCV*, 2011.
[5] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.
[6] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with ISA. In *CVPR*, 2011.
[7] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, 2004.
[8] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action Recognition by Dense Trajectories. In *CVPR*, 2011.
[9] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.


michael.sapienza-2011@brookes.ac.uk