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

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 perfackormed 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.

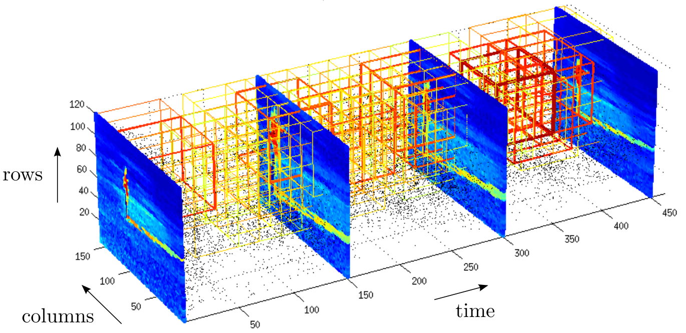

Figure 1: *A boxing video sequence taken from the KTH dataset [2] plotted in space and time. Overlaid on the video are discriminative cubic action subvolumes learned in a max-margin multiple instance learning framework [1], with colour indicating their class membership strength. Since the scene context of the KTH dataset is not discriminative of the particular action, only subvolumes around the actor were selected as positive instances.*

The contributions of this work are as follows:
i) We cast the conventionally supervised BoF action clip classification approach [3] into a weakly supervised setting, where clips are represented as bags of histogram instances with latent class variables. 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.
ii) In order to learn the subvolume class labels, we apply multiple instance learning (MIL) to 3D space-time videos, as we maintain that actions are better dackefined 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, 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.

The proposed action recognition system is composed of three main building blocks:
i) The description of space-time video blocks via histograms of Dense Trajectory features [4], which captures the trajectory's shape, appearance, and motion information.
ii) The representation of a video clip as a "bag of subvolumes" illustrated in Fig. 2, and the learning of positive subvolumes from weakly labelled training sequences within a max-margin MIL framework [1].
iii) The mapping of instance/subvolume scores to bag/clip scores by learning a hyperplane on instance margin features. Further details of the action recognition system are discussed in the methodology section of the paper.

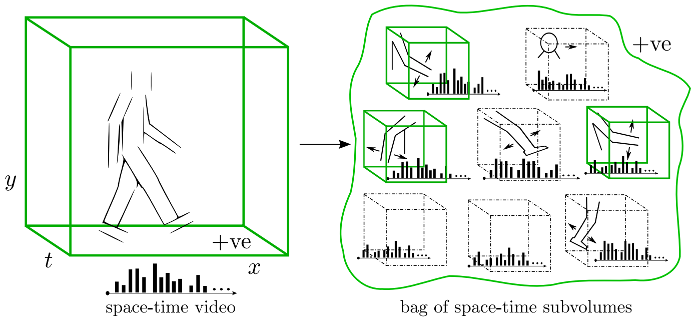

Figure 2: *Instead of defining an action as a space-time pattern in an entire video clip (left), we propose to define an action as a collection of space-time action parts contained in video subvolumes (right). The labels of each action subvolume are initially unknown. Multiple instance learning is used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

In order to validate our action recognition system, we evaluated its performance on ftour challenging action datasets, namely the **KTH** (6 classes), **YouTube** (11 classes), **Hollywood2** (12 classes) and **HMDB** (51 classes).

In 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*.

[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] 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.

[3] 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.

[4] 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.