

AMTnet: Action-Micro-Tube Regression by End-to-end Trainable Deep Architecture

Suman Saha Gurkirt Singh Fabio Cuzzolin
Oxford Brookes University, Oxford, United Kingdom

{suman.saha-2014, gurkirt.singh-2015, fabio.cuzzolin}@brookes.ac.uk

Abstract

Dominant approaches to action detection can only provide sub-optimal solutions to the problem, as they rely on seeking frame-level detections, to later compose them into ‘action tubes’ in a post-processing step. With this paper we radically depart from current practice, and take a first step towards the design and implementation of a deep network architecture able to classify and regress whole video subsets, so providing a truly optimal solution of the action detection problem. In this work, in particular, we propose a novel deep net framework able to regress and classify 3D region proposals spanning two successive video frames, whose core is an evolution of classical region proposal networks (RPNs). As such, our 3D-RPN net is able to effectively encode the temporal aspect of actions by purely exploiting appearance, as opposed to methods which heavily rely on expensive flow maps. The proposed model is end-to-end trainable and can be jointly optimised for action localisation and classification in a single step. At test time the network predicts ‘micro-tubes’ encompassing two successive frames, which are linked up into complete action tubes via a new algorithm which exploits the temporal encoding learned by the network and cuts computation time by 50%. Promising results on the J-HMDB-21 and UCF-101 action detection datasets show that our model does outperform the state-of-the-art when relying purely on appearance.

1. Introduction

In recent years most action detection frameworks [6, 33, 20, 22] employ deep convolutional neural network (CNN) architectures, mainly based on region proposal algorithms [29, 37, 21] and two-stream RGB and optical flow CNNs [24, 6]. These methods first construct training hypotheses by generating region proposals (or ‘regions of interest’, ROI¹), using either Selective Search [29], EdgeBoxes [37] or a region proposal network (RPN) [21]. ROIs are then sampled as positive and negative training examples as per the ground-truth. Subsequently, CNN features are ex-

tracted from each region proposal. Finally, ROI pooled features are fed to a softmax and a regression layer for action classification and bounding box regression, respectively.

This dominant paradigm for action detection [6, 33, 20, 22], however, only provides a *sub-optimal* solution to the problem. Indeed, rather than solving for

$$T^* \doteq \arg \max_{T \subset V} \text{score}(T), \quad (1)$$

where T is a subset of the input video of duration D associated with an instance of a known action class, they seek partial solutions for each video frame $R^*(t) \doteq \arg \max_{R \subset I(t)} \text{score}(R)$, to later compose in a post-processing step partial frame-level solutions into a solution $\hat{T} = [R^*(1), \dots, R^*(D)]$ of the original problem (1), typically called *action tubes* [6]. By definition, $\text{score}(\hat{T}) \leq \text{score}(T^*)$ and such methods are bound to provide suboptimal solutions. The post-processing step is essential as those CNNs do not learn the temporal associations between region proposals belonging to successive video frames. This way of training is mostly suitable for object detection, but inadequate for action detection where both spatial and temporal localisation are crucial. To compensate for this and learn the temporal dynamics of human actions, optical flow features are heavily exploited [6, 33, 20, 22].

With this paper we intend to initiate a research programme leading, in the medium term, to a new deep network architecture able to classify and regress whole video subsets. In such a network, the concepts of (video) region proposal and action tube will coincide.

In this work, in particular, we take a first step towards a truly optimal solution of the action detection problem by considering video region proposals formed by a pair of bounding boxes spanning two successive video frames at an arbitrary temporal interval Δ (see Figure 2). We call these pairs of bounding boxes *3D region proposals*. The advantages of this approach are that a) appearance features can be exploited to learn temporal dependencies (unlike what happens in current approaches), thus boosting detection performance; b) the linking of frame-level detections over time is no longer a post processing step and can be (partially) learned by the network. Obviously, at this stage we still need to construct action tubes from 3D region proposals.

¹A ROI is a rectangular bounding box parameterized as 4 coordinates in a 2D plane $[x1 \ y1 \ x2 \ y2]$.

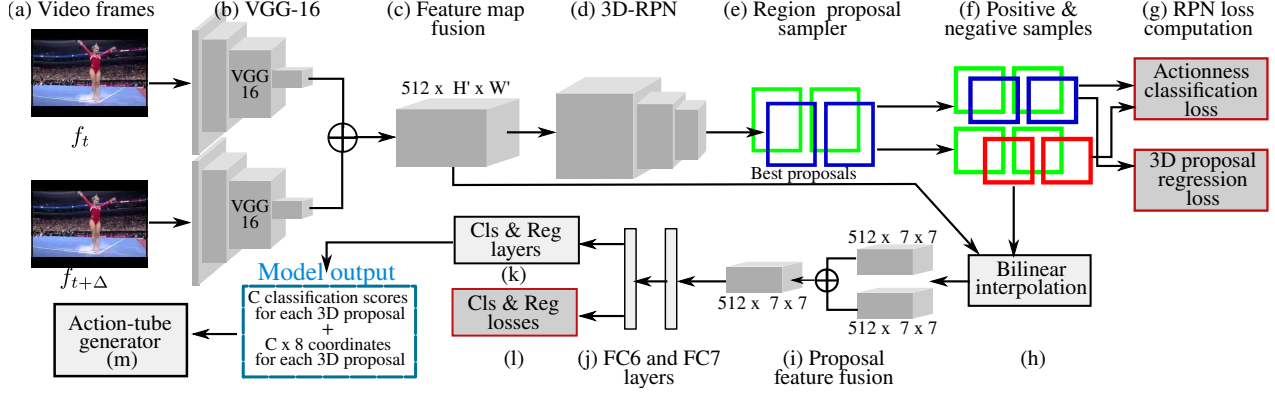


Figure 1. At train time, the input to the network is a pair of successive video frames (a) which are processed through two parallel VGG-16 networks (b). The feature maps generated by the last convolution layers are fused (c) and the fused feature map is fed to a 3D-RPN network (d). The RPN generates 3D region proposals and their associated actionness [2] scores which are then sampled as positive and negative training examples (f) by a proposal sampler (e). The sampled proposals and their scores are used to compute the actionness and 3D proposal regression losses (g). Subsequently, a bilinear feature pooling (h) and an element-wise feature fusion (i) are used to obtain a fixed sized feature representation for each sampled 3D proposal. Finally, the pooled and fused features are passed through fully connected (FC6 & FC7) (j), classification and regression (k) layers to train for action classification and a micro-tube regression. At test time, the predicted micro-tubes are linked in time by the action-tube generator (m).

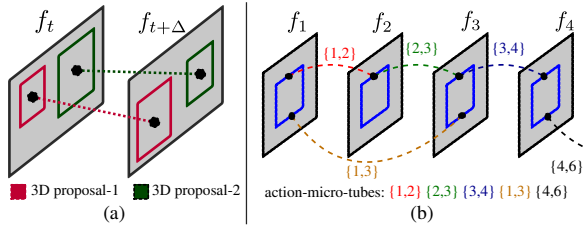


Figure 2. (a) The 3D region proposals generated by our 3D-RPN network span pairs of successive video frames f_t and $f_{t+\Delta}$ at temporal distance Δ . (b) Ground-truth action-micro-tubes generated from different pairs of successive video frames.

We thus propose a radically new approach to action detection based on (1) a novel deep learning architecture for regressing and classifying two-frame *micro-tubes*², illustrated in Figure 1, in combination with (2) an original strategy for linking micro-tubes up into proper action tubes. At test time, this new framework does not completely rely on post-processing for assembling frame-level detections, but makes use of the temporal encoding learned by the network. We show that: i) such a network trained on pairs of successive RGB video frames can learn the spatial and temporal extents of action instances relatively better than those trained on individual video frames, and ii) our model outperforms the current state-of-the-art [6, 33, 22] in spatio-temporal action detection by just exploiting appearance (the RGB video frames), in opposition to the methods which heavily exploit expensive optical flow maps.

Just to be clear, the aim of this paper is not to renounce to optical flow cues, but to move from frame-level detections

to whole tube regression. Indeed the method can be easily extended to incorporate motion at the micro-tube level rather than frame level, allowing fusion of appearance and motion at training time, unlike current methods [20, 22].

Overview of the approach. Our proposed network architecture (see Figure 1) employs and adapts some of the architectural components recently proposed in [21, 13].

At training time, the input to the model is a pair of successive video frames (a) which are fed to two parallel CNNs (b) (§ Section 3.1). The output feature maps of the two CNNs are fused (c) and passed as input to a 3D region proposal network (3D-RPN) (d) (§ Section 3.2). The 3D-RPN network generates 3D region proposals and their associated *actionness*³ [2] scores, which are then sampled as positive and negative training examples (f) by a proposal sampler (e) (§ 3.3). A training mini-batch of 256 examples are constructed from these positive and negative samples. The mini-batch is firstly used to compute the *actionness* classification and 3D proposal regression losses (g) (§ 4.1), and secondly, to pool CNN features (for each 3D proposal) using a bilinear interpolation layer (h) (§ 3.4).

In order to interface with the fully connected layers (j) (§ 3.5), bilinear interpolation is used to get a fixed-size feature representation for each variably sized 3D region proposal. As our 3D proposals consist of a pair of bounding boxes, we apply bilinear feature pooling independently on each bounding box in a pair, which gives rise to two fixed-size pooled feature maps of size $[512 \times kh \times kw]$, where $kh = kw = 7$ for each 3D proposal. We then apply element-wise fusion (i) (§ 3.4) to these 2 feature maps.

²We call ‘micro-tubes’ the 3D video region proposals, spanning pairs of successive frames, generated by the network at test time.

³The term *actionness* [2] is used to denote the possibility of an action being present within a 3D region proposal.

Each pooled and then fused feature map (representing a 3D proposal) is passed to two fully connected layers (FC6 and FC7) (**j**) (§ 3.5). The output of the FC7 layer is a fixed sized feature vector of shape $[4096 \times 1]$. These 4096 dimension feature vectors are then used by a classification and a regression layers (**k**) (§ 3.5) to output (**l**) $B \times C$ classification scores and (**2**) $B \times C \times 8$ coordinate values where B is the number of 3D proposals in a training mini-batch and C is the number of action categories in a given dataset.

At test time we select top 1000 predicted micro-tubes by using non-maximum suppression, modified to work with pairs of bounding boxes and pass these to an action-tube generator (**m**) (§ 5) which links those micro-tubes in time. At both training and test time, our model receives as input successive video frames $f_t, f_{t+\Delta}$. At training time we generate training pairs using 2 different Δ values 1 and 2 (§ 6.1). At test time we fix $\Delta = 1$. As we show in the supplementary material, even consecutive frames ($\Delta = 1$) carry significantly different information which affects the overall video-mAP. Throughout this paper, “3D region proposals” denotes the RPN-generated pairs of bounding boxes regressed by the middle layer (Figure 1 (g)), whereas “micro-tubes” refers to the 3D proposals regressed by the end layer (Figure 1 (l)).

Contributions. In summary, the key contributions of this work are: (**1**) on the methodological side, a key conceptual step forward from action detection paradigms relying on frame-level region proposals towards networks able to regress optimal solutions to the problem; (**2**) a novel, end-to-end trainable deep network architecture which addresses the spatiotemporal action localisation and classification task jointly using a single round of optimisation; (**3**) at the core of this architecture, a new design for a fully convolutional action localisation network (3D-RPN) which generates 3D video region proposals rather than frame-level ones; (**4**) a simple but efficient regression technique for regressing such 3D proposals; (**5**) a new action-tube generation algorithm suitable for connecting the micro-tubes so generated, which exploits the temporal encoding learnt by the network.

Experimental results on the J-HMDB-21 and UCF-101 action detection datasets show that our model outperforms state-of-the-art appearance-based models, while being competitive with methods using parallel appearance and flow streams. Finally, to the best of our knowledge, this is the first work in action detection which uses *bilinear interpolation* [8, 9] instead of the widely used RoI max-pooling layer [4], thus allowing gradients to flow backwards for both convolution features and coordinates of bounding boxes.

2. Related work

Deep learning architectures have been increasingly applied of late to action classification [12, 14, 24, 28], spatial [6], temporal [23] and spatio-temporal [33, 22, 20] ac-

tion localisation. While many works concern either spatial action localisation [18, 32, 10, 25] in trimmed videos or temporal localisation [17, 3, 27, 19, 31, 23, 35, 34] in untrimmed videos, only a handful number of methods have been proposed to tackle both problems jointly. Spatial action localisation has been mostly addressed using segmentation [18, 25, 10] or by linking frame-level region proposal [6, 33, 32]. Gkioxari and Malik [6], in particular, have built on [5] and [24] to tackle spatial action localisation in temporally trimmed videos, using Selective-Search [29] based region proposals on each frame of the videos.

Most recently, supervised frame-level action proposal generation and classification have been used by Saha *et al.* [22] and Peng *et al.* [20], via a Faster R-CNN [21] object detector, to generate frame level detections independently for each frame and link them in time in a post-processing step. Unlike [30, 6, 33], current methods [32, 22, 20] are able to leverage on end-to-end trainable deep-models [21] for frame level detection. However, tube construction is still tackled separately from region proposal generation.

Our novel network architecture, generates micro-tubes (the smallest possible video-level region proposals) which span across frames, and are labelled using a single soft-max score vector, in opposition to [6, 33, 20, 22] which generate frame-level region proposals. Unlike [6, 33, 20, 22], our proposed model is *end-to-end trainable* and requires a *single step of optimisation* per training iteration. To the contrary, [6, 33] use a multi-stage training strategy mutated from R-CNN object detection [5] which requires training two CNNs (appearance and optical-flow) independently, plus a battery of SVMs. The two most recent papers [20, 22] extend this Faster R-CNN [21] framework and train independently appearance and motion CNNs. Compared to [6, 33, 20, 22], which heavily exploit expensive optical flow maps, our model learns spatiotemporal feature encoding directly from raw RGB video frames.

3. Network Architecture

All the stages of Figure 1 are described below in detail.

3.1. Convolutional Neural Network

The convolutional (conv) layers of our network follow the VGG-16 architecture [24]. We use two parallel VGG-16 networks (§ Figure 1 (b)) to apply convolution over a pair of successive video frames. Each VGG-16 has 13 conv layers intermixed with 5 max pooling layers. Each conv layer has a 3×3 filter and 1×1 stride and padding. Each max pooling layer has filter shape 2×2 . We discard all the VGG-16 layers after the last (13-th) conv layer.

Feature map fusion. Our network takes two successive video frames f_t and $f_{t+\Delta}$ as inputs. For an input video frame of shape $[3 \times H \times W]$, the last conv layer of each VGG-16 outputs a feature map of shape $[D \times H' \times W']$ where

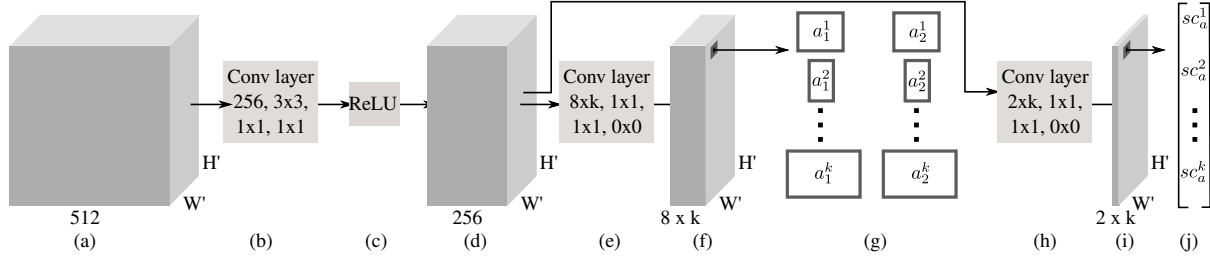


Figure 3. 3D-RPN architecture.

$D = 512$, $H' = \frac{H}{16}$, and $W' = \frac{W}{16}$. We fuse the two conv feature maps produced by the two parallel VGG-16 networks using element-wise sum fusion (§ Figure 1 (c)). As a consequence, the fused feature map encodes both appearance *and* motion information (for frames f_t and $f_{t+\Delta}$), which we pass as input to our 3D-RPN network.

Our new 3D region proposal network (Figure 1 (d)) builds on the basic RPN structure [21] to propose a fully convolutional network which can generate 3D region proposals via a number of significant architectural changes.

3.2. 3D region proposal network

3D region proposal generation. As we explained, unlike a classical RPN [21] which generates region proposals (rectangular bounding boxes) per image, our 3D-RPN network generates (video) region proposals spanning a pair of video frames. A single proposal thus consists of a pair of rectangular bounding boxes. The input to our 3D-RPN is a fused VGG-16 feature map (§ Figure 1 (c)) of size $[512 \times H' \times W']$. We generate anchor boxes in a similar way as in [21]: namely, we project back each point in the $H' \times W'$ grid (of the input feature map) onto the original image plane of size $H \times W$. For each projected point we generate k pairs of anchor boxes of different aspect ratios.

Let $(x_{a_i}, y_{a_i}, w_{a_i}, h_{a_i})$ denote the centroid, width and height of the anchor boxes in a pair. We use the subscript i to index the two boxes in a pair, i.e. $i = \{1, 2\}$. Similarly, $(x_{g_i}, y_{g_i}, w_{g_i}, h_{g_i})$ refer to the centroid, width and height of the ground-truth pair. We can transform a pair of input anchor boxes into a predicted pair of ground-truth boxes via⁴:

$$\begin{aligned} x_g &= x_a + \phi_x w_a & y_g &= y_a + \phi_y h_a \\ w_g &= w_a \exp(\phi_w) & h_g &= h_a \exp(\phi_h) \end{aligned} \quad (2)$$

where (ϕ_{x_i}, ϕ_{y_i}) specify a scale-invariant translation of the center of the anchor boxes, and (ϕ_{w_i}, ϕ_{h_i}) specify a log-space translation of their width and height.

Both RPN and the micro-tube regression layer (Figure 1 (k)) predict the bounding box regression offsets $(\phi_{x_i}, \phi_{y_i}, \phi_{w_i}, \phi_{h_i})$. Our anchor generation approach differs from that of [21], in the sense that we generate k pairs of anchors instead of k anchors.

⁴We removed the subscript i in Eq. 2 for sake of simplicity.

Network architecture. The network architecture of our 3D-RPN is depicted in Figure 3. To encode the location information of each pair of anchors, we pass the fused VGG-16 feature map through a 3×3 convolution (b), a rectified linear nonlinearity (c), and two more 1×1 convolution ((e) and (h)) layers. The first conv layer (b) consists of 256 convolution filters with 1×1 stride and padding, resulting in a feature map of size $[256 \times H' \times W']$ (d). The second conv layer (e) has $8 \times k$ convolution filters with 1×1 stride and does not have padding. It outputs a feature map of shape $[(8 \times k) \times H' \times W']$ (f) which encodes the location information (8 coordinate values) of $[k \times H' \times W']$ pairs of anchor boxes (g). The third conv layer (h) is the same as (e). The only difference is in the number of filters which is $2 \times k$ to encode the actionness score (i.e. probability of action or no-action) (j) for each k pairs of anchors.

As RPN is a fully convolutional neural network, classification and regression weights are learned directly from the convolution features, whereas in the fully connected layers (§ 3.5) we apply linear transformation layers for classification and regression. In our 3D-RPN, the convolution layer (e) is considered as the regression layer, as it outputs the 8 regression offsets per pair of anchor boxes; the convolution layer (h) is the classification layer.

3.3. 3D region proposal sampling

Processing all the resulting region proposals is very expensive. For example, with $k = 12$ and a feature map of size $[512 \times 38 \times 50]$, we get $12 \times 38 \times 50 = 22800$ pairs of anchor boxes. For this reason, we subsample them during both training and testing following the approach of [21] (§ Figure 1 (e)). We only make a slight modification in the sampling technique, as in our case one sample consists of a pair of bounding boxes, rather than a single box.

Training time sampling. During training, we compute the intersection over union (IoU) between a pair of ground-truth boxes $\{G_t, G_{t+\Delta}\}$ and a pair of proposal boxes $\{P_1, P_2\}$, so that, $\psi_1 = \text{IoU}(G_t, P_1)$ and $\psi_2 = \text{IoU}(G_{t+\Delta}, P_2)$. We consider $\{P_1, P_2\}$ as a positive example if $\psi_1 \geq 0.5$ and $\psi_2 \geq 0.5$, that is both IoU values are above 0.5. When enforcing this condition, there might be cases in which we do not have any positive pairs. To avoid such cases, we also consider as positive pairs those which

have maximal mean IoU $(\psi_1 + \psi_2)/2$ with the ground-truth pair. As negative examples we consider pairs for which both IoU values are below 0.3.

We construct a minibatch of size B in which we can have at most $B_p = B/2$ positive and $B_N = B - B_p$ negative training samples. We set $B = 256$. Note that the ground-truth boxes $\{G_t, G_{t+\Delta}\}$ in a pair belong to a same action instance but come from two different video frames $\{f_t, f_{t+\Delta}\}$. As there may be multiple action instances present, during sampling one needs to make sure that a pair of ground-truth boxes belongs to the same instance. To this purpose, we use the ground-truth tube-id provided in the datasets to keep track of instances.

Test time sampling. During testing, we use non-maximum suppression (NMS) to select the top $B = 1000$ proposal pairs. We made changes to the NMS algorithm to select the top B pairs of boxes based on their confidence. In NMS, one first selects the box with the highest confidence, to then compute the IoU between the selected box and the rest. In our modified version (i) we first select the pair of detection boxes with the highest confidence; (ii) we then compute the mean IoU between the selected pair and the remaining pairs, and finally (iii) remove from the detection list pairs whose IoU is above an overlap threshold th_{nms} .

3.4. Bilinear Interpolation

The sampled 3D region proposals are of different sizes and aspect ratios. We use *bilinear interpolation* [8, 9] to provide a fixed-size feature representation for them, necessary to pass the feature map of each 3D region proposal to the fully connected layer fc6 of VGG-16 (§ Figure 1 (j)), which indeed requires a fixed-size feature map as input.

Whereas recent action detection methods [20, 22] use max-pooling of region of interest (RoI) features which only backpropagates the gradients w.r.t. convolutional features, bilinear interpolation allows us to backpropagate gradients with respect to both (a) convolutional features and (b) 3D RoI coordinates. Further, whereas [20, 22] train appearance and motion streams independently, and perform fusion at test time, our model requires one-time training, and feature fusion is done at training time.

Feature fusion of 3D region proposals. As a 3D proposal consists of a pair of bounding boxes, we apply bilinear feature pooling independently to each bounding box in the pair. This yields two fixed-size pooled feature maps of size $[D \times kh \times kw]$ for each 3D proposal. We then apply element-wise sum fusion (§ Figure 1 (i)) to these 2 feature maps, producing an output feature map of size $[D \times kh \times kw]$. Each fused feature map encodes the appearance and motion information of (the portion of) an action instance which may be present within the corresponding 3D region proposal. In this work, we use $D = 512$, $kh = kw = 7$.

3.5. Fully connected layers

Our network employs two fully connected layers FC6 and FC7 (Figure 1 (j)), followed by an action classification layer and a micro-tube regression layer (Figure 1 (k)). The fused feature maps (§ Section 3.4) for each 3D proposal are flattened into a vector and passed through FC6 and FC7. Both layers use rectified linear units and dropout regularisation [13]. For each 3D region proposal, the FC7 layer outputs a 4096 dimension feature vector which encodes the appearance and motion features associated with the pair of bounding boxes. Finally, these 4096-dimensional feature vectors are passed to the classification and regression layers. The latter output $[B \times C]$ softmax scores and $[B \times C \times 8]$ bounding box regression offsets (§ 3.2), respectively, for B predicted micro-tubes and C action classes.

4. Network training

4.1. Multi-task loss function

As can be observed in Figures 1 and 3, our network contains two distinct classification layers.

The *mid classification layer* (§ Figure 3 (h)) predicts the probability p^m of a 3D proposal containing an action, $p^m = (p_0^m, p_1^m)$ over two classes (action vs. no action). We denote the associated loss by L_{cls}^m . The *end classification layer* (§ Figure 1 (k)) outputs a discrete probability distribution (per 3D proposal), $p^e = (p_0^e, \dots, p_C^e)$, over $C + 1$ action categories. We denote the associated loss as L_{cls}^e .

In the same way, the network has a mid (Figure 3 (e)) and an end (§ Figure 1 (k)) regression layer – the associated losses are denoted by L_{loc}^m and L_{loc}^e , respectively. Both regression layers output a pair of bounding box offsets ϕ^m and ϕ^e (cfr. Eq. 2). We adopt the parameterization of ϕ (§ 3.2) given in [5].

Now, each training 3D proposal is labelled with a ground-truth action class c^e and a ground-truth micro-tube (§ 1) regression target g^e . We can then use the multi-task loss [21]:

$$\begin{aligned} L(p^e, c^e, \phi^e, g^e, p^m, c^m, \phi^m, g^m) = & \\ & \lambda_{cls}^e L_{cls}^e(p^e, c^e) + \lambda_{loc}^e [c \geq 1] L_{loc}^e(\phi^e, g^e) + \\ & \lambda_{cls}^m L_{cls}^m(p^m, c^m) + \lambda_{loc}^m [c = 1] L_{loc}^m(\phi^m, g^m) \end{aligned} \quad (3)$$

on each labelled 3D proposal to jointly train for (i) action classification (p^e), (ii) micro-tube regression (ϕ^e), (iii) actionness classification (p^m), and (iv) 3D proposal regression (ϕ^m). Here, $L_{cls}^e(p^e, c^e)$ and $L_{cls}^m(p^m, c^m)$ are the cross-entropy losses for the true classes c^e and c^m respectively, where c^m is 1 if the 3D proposal is positive and 0 if it is negative, and $c^e = \{1, \dots, C\}$.

The second term $L_{loc}^e(\phi^e, g^e)$ is defined over an 8-dim

tuple of ground-truth micro-tube regression target coordinates: $g^e = (\{g_{x_1}^e, g_{y_1}^e, g_{w_1}^e, g_{h_1}^e\}, \{g_{x_2}^e, g_{y_2}^e, g_{w_2}^e, g_{h_2}^e\})$ and the corresponding predicted micro-tube tuple: $\phi^e = (\{\phi_{x_1}^e, \phi_{y_1}^e, \phi_{w_1}^e, \phi_{h_1}^e\}, \{\phi_{x_2}^e, \phi_{y_2}^e, \phi_{w_2}^e, \phi_{h_2}^e\})$. The fourth term $L_{loc}^m(\phi^m, g^m)$ is similarly defined over a tuple g^m of ground-truth 3D proposal regression target coordinates and the associated predicted tuple ϕ^m . The Iverson bracket indicator function $[c \geq 1]$ in (3) returns 1 when $c^e \geq 1$ and 0 otherwise; $[c = 1]$ returns 1 when $c^m = 1$ and 0 otherwise.

For both regression layers we use a smooth L1 loss in transformed coordinate space as suggested by [21]. The hyper-parameters λ_{cls}^e , λ_{loc}^e , λ_{cls}^m and λ_{loc}^m , in Eq. 3 weigh the relative importance of the four loss terms. In the following we set to 1 all four hyper-parameters.

4.2. Optimisation

We follow the end-to-end training strategy of [13] to train the entire network in a single optimisation step. We use stochastic gradient descent (SGD) to update the weights of the two VGG-16 convolutional networks, with a momentum of 0.9. To update the weights of other layers of the network, we use the Adam [15] optimiser, with parameter values $\beta_1 = 0.9$, $\beta_2 = 0.99$ and a learning rate of 1×10^{-6} . During the 1st training epoch, we freeze the weights of the convolution networks and update only the weights of the rest of the network. We start fine-tuning the layers of the two parallel CNNs after completion of 1st epoch. The first four layers of both CNNs are not fine-tuned for sake of efficiency. The VGG-16 pretrained ImageNet weights are used to initialise the convolutional nets. The rest of the network's weights are initialised using a Gaussian with $\sigma = 0.01$.

5. Action-tube generation

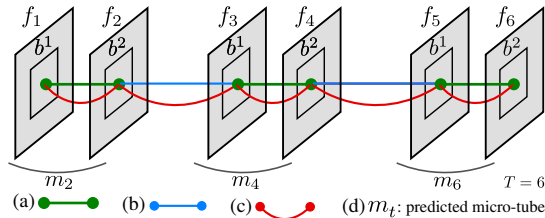


Figure 4. (a) The temporal associations learned by our network; (b) Our micro-tube linking algorithm requires $(T/2 - 1)$ connections; (c) the $T - 1$ connections required by [22]’s approach.

Once the predicted micro-tubes are regressed at test time, they need to be linked up to create complete action tubes associated with an action instance. To do this we introduce here a new action tube generation algorithm which is an evolution of that presented in [22]. There, temporally untrimmed action paths are first generated in a first pass of dynamic programming. In a second pass, paths are temporally trimmed to detect their start and end time. Here we

modify the first pass of [22] and build action paths using the temporal associations learned by our network. We use the second pass without any modification.

Linking up micro tubes (§ Figure 4) is not the same as linking up frame-level detections as in [22]. In the Viterbi forward pass of [22], the edge scores between bounding boxes belonging to consecutive video frames (i.e., frame f_t and f_{t+1}) are first computed. Subsequently, a DP (dynamic programming) matrix is constructed to keep track of the box indices with maximum edge scores. In the Viterbi backward pass, all consecutive pairs of frames, i.e. frames $\{1, 2\}, \{2, 3\}, \dots$ are traversed to join detections in time. Our linking algorithm saves 50% of the computing time, by generating edge scores between micro-tubes (which only needs $T/2 - 1$ iterations, cfr. Figure 4) rather than between boxes from consecutive frames (which, in the forward pass, needs $T - 1$ iterations). In the backward pass, the algorithm connects the micro-tubes as per the max edge scores.

Recall that a predicted micro-tube consists of a pair of bounding boxes (§ Figure 4), so that $m = \{b^1, b^2\}$. In the first pass action-specific paths $\mathbf{p}_c = \{m_t, t \in I = \{2, 4, \dots, T - 2\}\}$, spanning the entire video length are obtained by maximising via dynamic programming [6]:

$$E(\mathbf{p}_c) = \sum_{t \in I} s_c(m_t) + \lambda_o \sum_{t \in I} \psi_o(b^2_{m_t}, b^1_{m_{t+2}}), \quad (4)$$

where $s_c(m_t)$ denotes the softmax score (§ 3.5) of the predicted micro-tube m at time step t , the overlap potential $\psi_o(b^2_{m_t}, b^1_{m_{t+2}})$ is the IoU between the second detection box $b^2_{m_t}$ which forms micro-tube m_t and the first detection box $b^1_{m_{t+2}}$ of micro-tube m_{t+2} . Finally, λ_o is a scalar parameter weighting the relative importance of the pairwise term. By recursively removing the detection micro-tubes associated with the current optimal path and maximising (4) for the remaining micro-tubes we can account for multiple co-occurring instances of the same action class.

6. Experiments

6.1. Experimental setting

Datasets. All the experiments are conducted using the following two widely used action detection datasets: a) J-HMDB-21 [11] and b) UCF-101 24-class [26].

J-HMDB-21 is a subset of the relatively larger action classification dataset HMDB-51 [16], and is specifically designed for spatial action detection. It consists of 928 video sequences and 21 different action categories. All video sequences are temporally trimmed as per the action’s duration, and each sequence contains only one action instance. Video duration varies from 15 to 40 frames. Ground-truth bounding boxes for human silhouettes are provided for all 21 classes, and the dataset is divided into 3 train and test

splits. For evaluation on J-HMDB-21 we average our results over the 3 splits.

The **UCF-101** 24-class action detection dataset is a subset of the larger UCF-101 action classification dataset, and comprises 24 action categories and 3207 videos for which spatiotemporal ground-truth annotations are provided. We conduct all our experiments using the first split. Compared to J-HMDB-21, the UCF-101 videos are relatively longer and temporally untrimmed, i.e., action detection is to be performed in both space and time. Video duration ranges between 100 and 1000 video frames.

Note that the THUMOS [7] and ActivityNet [1] datasets are not suitable for spatiotemporal localisation, as they lack bounding box annotation.

Evaluation metrics. As evaluation metrics we use both: (1) *frame-AP* (the average precision of detections at the frame level) as in [6, 20]; (2) *video-AP* (the average precision of detection at video level) as in [6, 33, 22, 20]. We select an IoU threshold (δ) range [0.1:0.1:0.5] for J-HMDB-21 and [0.1,0.2,0.3] for UCF-101 when computing video-mAP. For frame-mAP evaluation we set $\delta = 0.5$.

Training data sampling strategy. As the input to our model is a pair of successive video frames and their associated ground-truth micro-tubes, training data needs to be passed in a different way than in the frame-level training approach [6, 33, 20, 22], where inputs are individual video frames. In our experiments, we use 3 different sampling schemes to construct training examples using different combinations of successive video frames (§ Figure 2 (b)): (1) *scheme-11* generates training examples from the pairs of frames $\{t=1, t=2\}$, $\{t=2, t=3\}$...; *scheme-21* uses the (non-overlapping) pairs $\{1, 2\}$, $\{3, 4\}$...; *scheme-32* constructs training samples from the pairs $\{1, 3\}$, $\{4, 6\}$...

In the supplementary material we explain all implementation details including data preprocessing, training/testing time requirements, training batch construction.

6.2. Model evaluation

We first show how a proper positive IoU threshold is essential during the sampling of 3D region proposals at training time (§ 3.3). Secondly, we assess whether our proposed network architecture, coupled with the new data sampling strategies (Sec. 6.1), improves detection performance. We then show that our model outperforms the appearance-based model of [22]. Finally, we compare the performance of the overall detection framework with the state-of-the-art.

Effect of different positive IoU thresholds on detection performance. We train our model on UCF-101 using two positive IoU thresholds: 0.7 and 0.5 (§ 3.3). The detection results (video-mAP) of these two models (Model-0-7 & -0-5) are shown in Table 1. Whereas [21] recommends an IoU threshold of 0.7 to subsample positive region proposals during training, in our case we observe that an IoU

Table 1. Effect of different positive IoU thresholds on detection performance (video-mAP).

IoU threshold δ	0.1	0.2	0.3
Model-0-7	64.04	54.83	44.664
Model-0-5	68.85	60.06	49.78

threshold of 0.5 works better with our model. Indeed, during sampling we compute IoUs between pairs of bounding boxes and then take the mean IoU to subsample (§ 3.3). As the ground-truth boxes (micro-tubes) are connected in time and span different frames, it is harder to get enough positive examples with a higher threshold like 0.7. Therefore, in the remainder we use an IoU of 0.5 for evaluation.

Effect of our training data sampling strategy on detection performance. *JHMDB-21 frame-mAP.* We first generate a J-HMDB-21 training set using the *scheme-11* (§ 6.1) and train our model. We then generate another training set using *scheme-32*, and train our model on the combined training set (*set-11+32*). Table 2 shows the per class frame-AP obtained using these two models. We can observe that out of 21 JHMDB action classes, the frame-APs of 15 classes actually improve when training the model on the new combined trainset (*set-11+32*). Overall performance increases by 1.64%, indicating that the network learns temporal association more efficiently when it is trained on pairs generated from different combinations of successive video frames.

JHMDB-21 video-mAP. The two above trained models are denoted by *Model-11* and *Model-11+32* in Table 4, where the video-mAPs at different IoU threshold for these two models are shown. Although the first training strategy *scheme-11* already makes use of all the video frames present in J-HMDB-21 training splits, when training our model using the combined trainset we observe an improvement in the video-mAP of 1.04% at $\delta = 0.5$.

Effect of exploiting appearance features. Further, we show that our model exploits appearance features (raw RGB frames) efficiently, contributing to an improvement of video-mAP by 3.2% over [22]. We generate a training set for UCF-101 split 1 using the training *scheme-21* and compare our model’s performance with that of the appearance-based model (*A) of [22]. We show the comparison in Table 3.

Note that, among the 24 UCF-101 action classes, our model exhibits better video-APs for 14 classes, with an overall gain of 3.2%. We can observe that, although trained on appearance features only, our model improves the video-APs significantly for action classes which exhibit a large variability in appearance and motion. Also, our model achieves relatively better spatiotemporal detection on action classes associated with video sequences which are significantly temporally untrimmed, such as *BasketballDunk*, *GolfSwing*, *Diving* with relative video-AP improvements of

Table 2. Effect of our training data sampling strategy on per class frame-AP at IoU threshold $\delta = 0.5$, JHMDB-21 (averaged over 3 splits).

frame-AP(%)	brushHair	catch	clap	climbStairs	golf	jump	kickBall	pick	pour	pullup	push	run	shootBall	shootBow	shootGun	sit	stand	swingBaseball	throw	walk	wave	mAP
ours (*)	46.4	40.7	31.9	62.3	91.0	4.3	17.3	29.5	86.2	82.7	66.9	35.5	33.9	78.2	49.7	11.7	13.8	57.1	21.3	27.8	27.1	43.6
ours (**)	43.7	43.6	33.0	61.5	91.8	5.6	23.8	31.5	91.8	84.1	73.1	32.3	33.3	81.4	55.1	12.4	14.7	56.3	22.2	24.7	29.4	45.0
Improvement	-2.6	2.9	1.0	-0.8	0.7	1.2	6.4	1.9	5.5	1.4	6.1	-3.2	-0.6	3.2	5.4	0.6	0.8	-0.8	0.8	-3.1	2.3	1.4
[6]	65.2	18.3	38.1	39.0	79.4	7.3	9.4	25.2	80.2	82.8	33.6	11.6	5.6	66.8	27.0	32.1	34.2	33.6	15.5	34.0	21.9	36.2
[32]	60.1	34.2	56.4	38.9	83.1	10.8	24.5	38.5	71.5	67.5	21.3	19.8	11.6	78.0	50.6	10.9	43.0	48.9	26.5	25.2	15.8	39.9
[33]	73.3	34.0	40.8	56.8	93.9	5.9	13.8	38.5	88.1	89.4	60.5	21.1	23.9	85.6	37.8	34.9	49.2	36.7	16.8	40.5	20.5	45.8
[20]	75.8	38.4	62.2	62.4	99.6	12.7	35.1	57.8	96.8	97.3	79.6	38.1	52.8	90.8	62.7	33.6	48.9	62.2	25.6	59.7	37.1	58.5
video-AP(%)																						
ours (*)	53.9	54.4	39.8	68.2	96.1	5.69	39.6	34.9	97.1	93.5	84.1	53.7	43.6	93.2	64.5	20.9	22.8	72.1	23.2	39.4	37.8	54.27
ours (**)	51.9	54.5	41.2	66.6	94.8	7.8	48.7	33.7	97.6	92.5	87.6	49.0	37.4	92.7	75.8	21.6	27.1	73.3	24.3	37.7	44.7	55.31
Improvement	-1.9	0.01	1.4	-1.6	-1.2	2.1	9.1	-1.2	0.4	-1.0	3.4	-4.7	-6.2	-0.5	11.2	0.6	4.2	1.1	1.1	-1.7	6.8	1.04
[6]	79.1	33.4	53.9	60.3	99.3	18.4	26.2	42.0	92.8	98.1	29.6	24.6	13.7	92.9	42.3	67.2	57.6	66.5	27.9	58.9	35.8	53.3
[32]	76.4	49.7	80.3	43.0	92.5	24.2	57.7	70.5	78.7	77.2	31.7	35.7	27.0	88.8	76.9	29.8	68.6	72.8	31.5	44.4	26.2	56.4

*Model-11 **Model-11+32

Table 3. Per class video-AP comparison at IoU threshold $\delta = 0.2$, UCF-101.

video-AP(%)	BasketballDunk	Biking	Diving	Fencing	FloorGymnastics	GolfSwing	IceDancing	LongJump	PoleVault	RopeClimbing	Skiing	Skijet	SoccerJuggling	WalkingWithDog	mAP
[22] (*A)	22.7	56.1	89.7	86.9	93.8	59.9	59.2	41.5	48.9	77.8	68.4	88	34.6	73.3	56.86
ours	39.6	59.5	91.2	88.5	94.1	70.7	70.4	49.8	71.0	97.2	74.0	92.9	80.2	73.6	60.06
Improvement	16.9	3.4	1.5	1.6	0.3	10.8	11.2	8.3	22.1	19.4	5.6	4.9	45.6	0.3	3.2

*A: appearance model

Table 4. Effect of our training data sampling strategy on video-mAP, JHMDB-21 (averaged over 3 splits).

IoU threshold δ	0.1	0.2	0.3	0.4	0.5
Model-11	57.73	57.70	57.60	56.81	54.27
Model-11+32	57.79	57.76	57.68	56.79	55.31

Table 5. Spatio-temporal action detection performance (video-mAP) comparison with the state-of-the-art on J-HMDB-21.

IoU threshold δ	0.1	0.2	0.3	0.4	0.5
Gkioxari and Malik [6]	-	-	-	-	53.30
Wang <i>et al.</i> [32]	-	-	-	-	56.40
Weinzaepfel <i>et al.</i> [33]	-	63.1	-	-	60.70
Saha <i>et al.</i> [22] (Spatial Model)	52.99	52.94	52.57	52.22	51.34
Peng and Schmid [20]	-	74.3	-	-	73.1
Ours	57.79	57.76	57.68	56.79	55.31

Table 6. Spatio-temporal action detection performance (video-mAP) comparison with the state-of-the-art on UCF-101.

IoU threshold δ	0.1	0.2	0.3	0.5	0.75	0.5:0.95
Yu <i>et al.</i> [36]	42.8	26.50	14.6	-	-	-
Weinzaepfel <i>et al.</i> [33]	51.7	46.8	37.8	-	-	-
Peng and Schmid [20]	77.31	72.86	65.70	30.87	01.01	07.11
Saha <i>et al.</i> [22] (*A)	65.45	56.55	48.52	-	-	-
Saha <i>et al.</i> [22] (full)	76.12	66.36	54.93	-	-	-
Ours - ML	68.85	60.06	49.78	-	-	-
Ours - ML - (*)	70.71	61.36	50.44	32.01	0.4	9.68
Ours - 2PDP - (*)	71.3	63.06	51.57	33.06	0.52	10.72

(*) cross validated alphas as in [22]; 2PDP - tube generation algorithm [22]

ML - our micro-tube linking algorithm.

16.9%, 10.8% and 1.5% respectively. We report significant gains in absolute video-AP for action categories *SoccerJuggling*, *PoleVault*, *RopeClimbing*, *BasketballDunk*, *IceDancing*, *GolfSwing* and *LongJump* of 45.6%, 22.1%, 19.4%, 16.9%, 11.2% 10.8% and 8.3%, respectively.

Detection performance comparison with the state-of-the-art. Table 5 reports action detection results, averaged over the three splits of *J-HMDB-21*, and compares them with those to our closest competitors. Note that, although our model only trained using the appearance features (RGB images), it outperforms [6] which was trained using both

appearance and optical flow features. Also, our model outperforms [22]’s spatial detection network.

Table 6 compares the action detection performance of our model on the UCF-101 dataset to that of current state of the art approaches. We can observe that our model outperforms [36, 33, 22] by a large margin. In particular, our appearance-based model outperforms [33] which exploits both appearance and flow features. Also notice, our method works better than that of [20] at higher IoU threshold, which is more useful in real-world applications.

7. Conclusions

In this work we departed from current practice in action detection to take a step towards deep network architectures able to classify and regress whole video subsets. In particular, we propose a novel deep net framework able to regress and classify 3D region proposals spanning two successive video frames, effectively encoding the temporal aspect of actions using just raw RGB values. The proposed model is end-to-end trainable and can be jointly optimised for action localisation and classification using a single step of optimisation. At test time the network predicts ‘micro-tubes’ spanning two frames, which are linked up into complete action tubes via a new algorithm of our design. Promising results confirm that our model does indeed outperform the state-of-the-art when relying purely on appearance.

Much work will need to follow. It remains to be tested whether optical flow can be integrated in this framework and further boost performance. As the search space of 3D proposals is twice the dimension of that for 2D proposals, efficient parallelisation and search are crucial to fully exploit the potential of this approach. Further down the road we wish to extend the idea of micro-tubes to longer time intervals, posing severe challenges in terms of efficient regression in higher-dimensional spaces.

References

- [1] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 7
- [2] W. Chen, C. Xiong, R. Xu, and J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 748–755, 2014. 2
- [3] A. Gaidon, Z. Harchaoui, and C. Schmid. Temporal localization of actions with actoms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2782–2795, 2013. 3
- [4] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 3
- [5] R. Girshick, J. Donahue, T. Darrel, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2014. 3, 5
- [6] G. Gkioxari and J. Malik. Finding action tubes. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2015. 1, 2, 3, 6, 7, 8
- [7] A. Gorban, H. Idrees, Y. Jiang, A. R. Zamir, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2015. 7
- [8] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 3, 5
- [9] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 3, 5
- [10] S. D. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In *Computer Vision–ECCV 2014*, pages 656–671. Springer, 2014. 3
- [11] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. Black. Towards understanding action recognition. 2013. 6
- [12] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231, Jan 2013. 3
- [13] J. Johnson, A. Karpathy, and L. Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 5, 6
- [14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2014. 3
- [15] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [16] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011. 6
- [17] I. Laptev and P. Pérez. Retrieving actions in movies. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 3
- [18] J. Lu, r. Xu, and J. J. Corso. Human action segmentation with hierarchical supervoxel consistency. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, June 2015. 3
- [19] D. Oneata, J. Verbeek, and C. Schmid. Efficient action localization with approximately normalized fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2545–2552, 2014. 3
- [20] X. Peng and C. Schmid. Multi-region two-stream R-CNN for action detection. In *ECCV 2016 - European Conference on Computer Vision*, Amsterdam, Netherlands, Oct. 2016. 1, 2, 3, 5, 7, 8
- [21] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 1, 2, 3, 4, 5, 6, 7
- [22] S. Saha, G. Singh, M. Sapienza, P. H. S. Torr, and F. Cuzolin. Deep learning for detecting multiple space-time action tubes in videos. In *British Machine Vision Conference*, 2016. 1, 2, 3, 5, 6, 7, 8
- [23] Z. Shou, D. Wang, and S. Chang. Action temporal localization in untrimmed videos via multi-stage cnns. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, June 2016. 3
- [24] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc., 2014. 1, 3
- [25] K. Soomro, H. Idrees, and M. Shah. Action localization in videos through context walk. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 3
- [26] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human action classes from videos in the wild. Technical report, CRCV-TR-12-01, 2012. 6
- [27] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2642–2649. IEEE, 2013. 3
- [28] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *Proc. Int. Conf. Computer Vision*, 2015. 3
- [29] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *Int. Journal of Computer Vision*, 2013. 1, 3
- [30] J. C. van Gemert, M. Jain, E. Gati, and C. G. Snoek. APT: Action localization proposals from dense trajectories. In *BMVC*, volume 2, page 4, 2015. 3
- [31] H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, pages 1–20, 2015. 3
- [32] L. Wang, Y. Qiao, X. Tang, and L. Van Gool. Actionness estimation using hybrid fully convolutional networks. *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2016. 3, 8

- [33] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, June 2015. [1](#), [2](#), [3](#), [7](#), [8](#)
- [34] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *arXiv preprint arXiv:1507.05738*, 2015. [3](#)
- [35] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. *CVPR*, 2016. [3](#)
- [36] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1302–1311, 2015. [8](#)
- [37] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014*, pages 391–405. Springer, 2014. [1](#)