Action localization with tubelets from motion

Mihir Jain[†⋆] Jan van Gemert[⋆] Hervé Jégou[†] Patrick Bouthemy[†] Cees G.M. Snoek[⋆]

[†]Inria [⋆]University of Amsterdam

Abstract

This paper considers the problem of action localization, where the objective is to determine when and where certain actions appear. We introduce a sampling strategy to produce 2D+t sequences of bounding boxes, called tubelets. Compared to state-of-the-art alternatives, this drastically reduces the number of hypotheses that are likely to include the action of interest. Our method is inspired by a recent technique introduced in the context of image localization. Beyond considering this technique for the first time for videos, we revisit this strategy for 2D+t sequences obtained from super-voxels. Our sampling strategy advantageously exploits a criterion that reflects how action related motion deviates from background motion.

We demonstrate the interest of our approach by extensive experiments on two public datasets: UCF Sports and MSR-II. Our approach significantly outperforms the state-of-the-art on both datasets, while restricting the search of actions to a fraction of possible bounding box sequences.

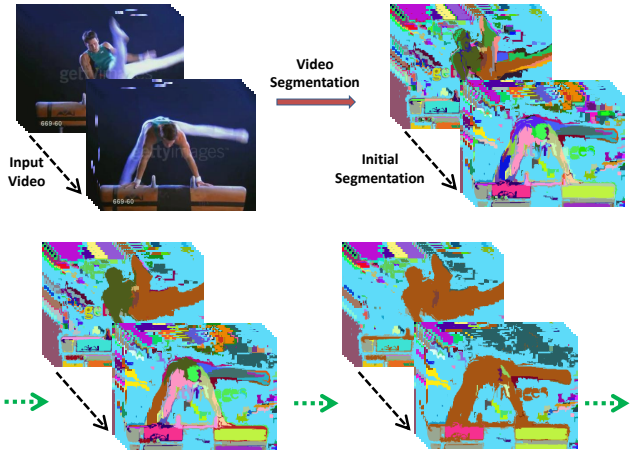

Figure 1. Overview of tubelets from motion: From an initial spatio-temporal segmentation in super-voxels, such as the one we propose based on motion, we produce additional super-voxels by merging them based on a criterion capturing the motion similarity. This produces a small set of tubelets, which is fed to a classifier.

1. Introduction

Recognizing actions in videos is an active area of research in computer vision. Because of the many fine-grained spatio-temporal variations in action appearance, the current performance is far from that achieved in other recognition tasks such as image search. The goal of action classification is to determine *which* action appears in the video. Temporal action detection estimates, additionally, *when* it occurs. This paper specifically considers the problem of action localization: the objective is to detect when *and* where an action of interest occurs. The expected output of such an action localization system is typically a subvolume encompassing the action of interest. Since a localized action only covers a fraction of the spatio-temporal volume in a video, the task is considerably more challenging than action classification and temporal detection. This task can be seen as the video counterpart of object detection in still images.

There is a large body of literature that aims at bypassing the costly sliding window approach [31]. The general strategy is to limit the set of tested windows to an acceptable number by varying optimization strategies such as efficient sub-window search [15] (branch and bound search), objectness [2] and, more recently, a "selective search" strategy [29]. The latter generates a set of category-independent candidate windows by iteratively agglomerating super-pixels based on several similarity criteria. It achieves, on average, a similar accuracy as that obtained by Deformable Part Models [10] (DPM), while drastically reducing the number of box hypotheses to be tested.

Most action localization systems are inspired by the aforementioned object detection strategies. For instance, Yuan *et al*. have extended the branch and bound approach to videos [36], while Tian *et al*. [24] have proposed spatio-temporal DPM (SDPM). A noticeable exception is selective search [29]: To the best of our knowledge and despite its amenability to handle varying aspect ratios (in this respect, better than DPM), it has never been explored for videos.

Our first contribution is to investigate the selective search sampling strategy for videos. We adopt the general principle and extend it. First, we consider super-voxels instead of super-pixels to produce spatio-temporal shapes. This directly gives us $2D+t$ sequences of bounding boxes, referred to as *tubelets* in this paper, without the need to address the

problem of linking boxes from one frame to another, as required in other approaches [25, 26].

Our second contribution is explicitly incorporating motion information in various stages of the analysis. We introduce *independent motion evidence* as a feature to characterize how the action motion deviates from the background motion. By analogy to image descriptors such as the Fisher vector [19], we encode the singularity of the motion in a feature vector associated with each super-voxel. First, motion is used as a merging criterion in the agglomerative stage of our sampling strategy. Second, motion is used as an independent cue to produce super-voxels partitioning the video.

Our approach offers several advantages. We produce a small set of candidate tubelets, which allows us to describe each tubelet with a high-quality, computationally expensive representation. Furthermore, the bounding boxes are tailored to super-voxel shapes, which tend to improve the spatio-temporal localization of our bounding box sequences. As a result, we observe a consistent and significant gain over concurrent approaches for action localization. This is not surprising, as the still image object detection counterpart was recently shown to outperform DPM, as demonstrated in the VOC 2012 challenge [8]. Our motion-based adaptation brings a large benefit, as shown by comparing with more naive motion-free adaptation of selective search to videos.

2. Related work

In this section, we present existing works into more details, in order to position our method with respect to the literature. Most references address recognition tasks in videos, but our work is also related to papers on object recognition in images, in particular object localization.

Action recognition and localization. Current action recognition methods determine which action occurs in a video with good accuracy [9, 13, 23, 30, 32]. The task of localization is more demanding as it also requires specifying the location where the action happens in the video. This location is often expressed as a cuboid referred to as subvolume [4, 24, 36]. Subvolume-based detection is inadequate in the case of complex actions, when the actor moves spatially or when the aspect ratio varies significantly like exemplified in Figure 2. Recently, action location is more precisely defined as a sequence of bounding boxes [16, 26, 27]. The corresponding 2D+t volume, which we refer to as tubelet, tightly bounds the actions in the video space and provides a more accurate spatio-temporal localization of actions. However, the methods considering this definition are more costly since the search space is significantly larger [26] than in subvolume-based localization. Therefore, it is critical to have a high-quality sub-sampling strategy for tubelets, as we propose in our paper.

We have recently witnessed a trend for methods aiming at providing a more precise localization, for instance for obtaining generic spatio-temporal human tracks [14] using a human detector-tracker. In another work [16], the detector and tracker are avoided by treating the actor location as a latent variable. Raptis *et al.* [21] select trajectory groups that serve as candidates for the parts of an action. While such a mid-level representation assists recognition, they localize only parts of actions. In [25, 26], candidate bounding boxes are generated for each frame separately and then the optimal spatio-temporal path is found by Max-Path Search. However, this approach uses a sliding-window object detector, which is not only impractical on large video datasets, but it is also unsuitable for actions with varying aspect ratios. Rather than considering a video as a set of images and finding optimal spatio-temporal path later, we prefer to consider it as a spatio-temporal source from the very beginning.

Trichet and Nevatia [28] propose spatio-temporal tubes for video segmentation. The only method we are aware of that uses tubelet-like representation for action localization is by Wang *et al*. [33], that appeared in the meantime. They model human action as a spatio-temporal tube of maximum mutual information of feature trajectories towards the action class. One of the advantages we have over them is that our approach produces class-independent hypotheses.

Extensions from object localization. Many action localization approaches are inspired by box sampling strategies adapted from the object detection literature in still images. The most popular is the sliding-window approach, extended to sliding-subvolume for actions. Due to its considerable computational cost in object localization, not to mention in videos, many works have attempted to circumvent sliding windows such as efficient sub-window search [15].

Rather than reducing the number of sliding windows, category-independent object proposals have been proposed to aid object localization [1, 7, 17, 20]. The object proposals produced by these methods are 2D-locations likely to contain any object. This class of approaches was shown successful for salient object detection [11], weakly supervised object localization [6], and supervised object detection [29].

In our paper, the goal is to generate flexible tubelets that are independent of the action category. Our approach is inspired by the object sampling of selective search [29], yet specifically considers the spatio-temporal context of video localization. In this context and as shown in our paper, motion is a key feature and our method explicitly takes the motion into account when generating tubelets. Since actions are highly non-rigid, we use a flexible over-segmentation of the video into super-voxels. Super-voxels give excellent boundary recall [3, 34, 35] for non-rigid objects. Thus, in analogy to the 2D super-pixel methods used for static object proposals [1, 17, 29], we use super-voxels as the main mechanism to build video tubelets.

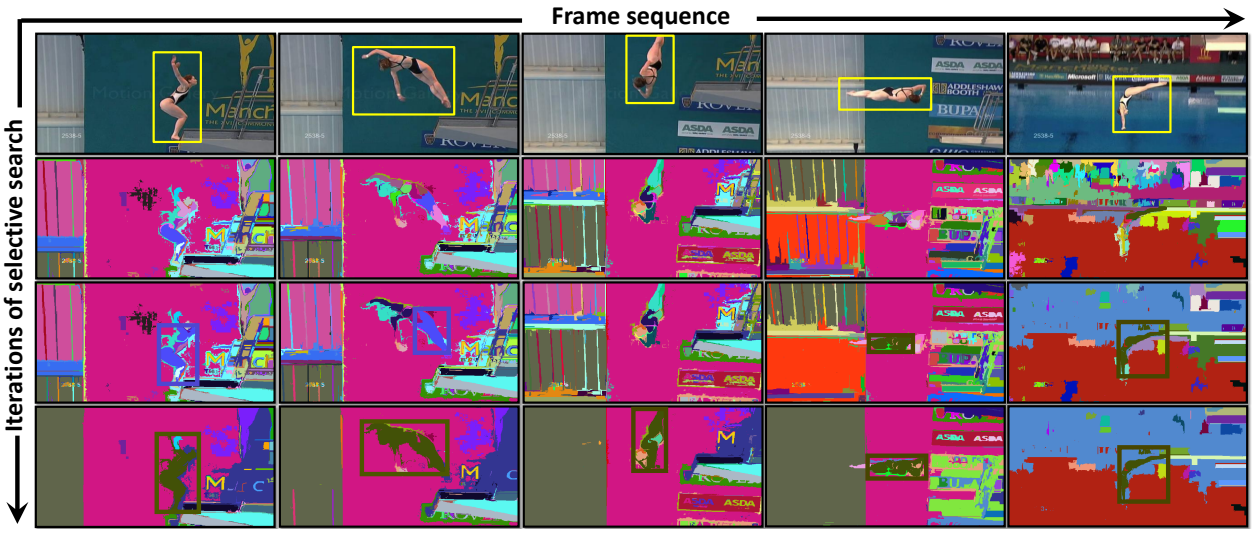

Figure 2. Illustration of hierarchical sampling of tubelets. *Top.* A sampled sequence of frames (1^{st}, 15^{th}, 25^{th}, 35^{th}, 50^{th}) associated with the action 'diving' from UCF-Sports dataset. The yellow bounding boxes represent the ground-truth tubelet. *Row 2* shows the video segmentation used as input to our method. *The last two rows* show two stages of the hierarchical grouping algorithm. A tubelet close to the action is also represented by bounding boxes in each row. Observe how it is close to the ground-truth tubelet in the last row.

3. Action sequence hypotheses: Tubelets

This section describes our approach for iteratively sampling a set of candidate box sequences or *tubelets*. We generalize the selective search [29] method from images to videos to delineate spatio-temporal action sequences. This generalization from $2D$ to $2D + t$ demands adaptation to the characteristics of video, such as relying on super-voxels instead of super-pixels.

We first give a brief overview of the action localization pipeline. Then we describe how tubelets are sampled iteratively. Finally, we focus on an important aspect of the technique, *i.e.*, the merging criteria and the video features upon which they are built. Later in Section 4, we further extend this approach by incorporating motion in two stages of the processing pipeline.

3.1. Overview of the action localization pipeline

1. **Super-voxel segmentation.** To generate the initial set of super-voxels, we first rely on a third-party Graph-based (GB) video segmentation method [34]. We choose GB over other segmentation methods in [34] because it is more efficient w.r.t. time and memory, *i.e.*, about 13 times faster than a slightly more accurate hierarchical version (GBH) [34]. This step produces n super-voxels, to which we associate n tubelets, obtained as the sequences of bounding boxes that tightly encompass the super-voxels.
2. **Iterative generation of additional tubelets.** It consists of $n-1$ iterations. Each merges two super-voxels into a new one. The choice of the two super-voxels to be merged in a given iteration depends on a similarity criterion that we will detail in the following subsections.
3. **Descriptor computation.** This step computes a bag-of-words (BOW) representation for each tubelet. As local descriptor we employ MBH [5].
4. **Classification step.** BOW histograms of tubelets are used for training a classifier per class.

3.2. Hierarchical sampling of tubelets

In this section, our objective is to produce additional tubelets from successive merging of the super-voxels produced by the initial spatio-temporal segmentation. The algorithm is inspired by the selective search for object localization in images [29].

Super-voxel generation. We iteratively merge super-voxels in an agglomerative manner. Starting from the initial set of super-voxels, we hierarchically group voxels until the video becomes a single super-voxel. At each iteration, a new super-voxel is produced from two super-voxels, which are no longer considered in subsequent iterations.

Formally, we produce a hierarchy of super-voxels that are represented as a tree: the leaves correspond to the initial super-voxels while the internal nodes are produced by the merge operations. The root node is the whole video and the corresponding super-voxel is produced in the last iteration. Since this hierarchy of super-voxels is organized as a binary tree, it is straightforward to show the algorithm produces $n-1$ additional super-voxels.

Tubelets. In each frame where a super-voxel appears it is tightly bounded by a bounding box rectangle. A sequence of frames with such bounding boxes forms a tubelet. The

Figure 3. Example from action 'Running': The first two images depict a video frame and the initial super-voxel segmentation used as input of our approach. The four other images represent the segmentation after a varying number of merge operations.

hierarchical algorithm samples tubelets with spatial boxes at all scales and sequences of all possible lengths in time. Note that a tubelet is a more general shape than the cuboids [16, 26, 27]. As the output of the algorithm, we have $2n-1$ tubelets, $n-1$ obtained from the new super-voxels and n from the segmentation.

The merge starts by selecting the two super-voxels to be merged. For this purpose, we rely on similarities computed between all the neighboring super-voxels that are still active. The similarity measures are detailed in the next subsection. After the merge, we compute the new similarities between the resulting super-voxel and its neighbors.

Figure 2 illustrates the method on a sample video. Each color represents a super-voxel and after every iteration a new entry is added and two are removed. After 1000 iterations, observe that two tubelets (blue and dark green) emerge around the action of interest in the beginning and the end of the video, respectively. At iteration 1720, the two corresponding super-voxels are merged. The novel tubelet (dark green) resembles the ground truth yellow tubelet. This exhibits the ability of our method to group tubelets both spatially and temporally. As importantly, it shows the capability to sample a tubelet with boxes having very different aspect ratios. This is unlikely to be coped by sliding-subvolumes or even approaches based on efficient sub-window search.

Figure 3 depicts another example, with a single frame considered at different stages of the algorithm. Here the initial super-voxels (second image) are spatially more decomposed because the background is cluttered both in appearance and in motion (spectators cheering). Even in such a challenging case our method is able to group the super-voxels related to the action of interest.

3.3. Merging criteria: Similarity measures

We employ five complementary similarity measures to compare super-voxels, in order to select which ones should be merged. They are fast to compute. Four of these measures are adapted from selective search in still images [29] where super-pixels are used. We revise these measures based on color, texture, size and fill for super-voxels. In addition and because our objective is not to segment the objects but to delineate the action or actors, we additionally employ a motion-based similarity encoding *independent motion evidence* (IME) to characterize a super-voxel.

Merging with color, texture and motion: s_{C}, s_{T}, s_{M}. These three similarity measures are computed in a similar manner: They describe each super-voxel with a histogram and for comparison histogram intersection is used. They differ only in the way the histograms are computed from different characteristics of a given super-voxel:

- The color histogram h_{C} captures the HSV components of the pixels included in a super-voxel;
- h_{T} encodes the texture or gradient information of a given super-voxel;
- Our merging criterion is based on a histogram h_{M} computed from our IME feature, which is detailed in the section 4 devoted to motion.

As the process of merging is the same for each of the histogram representations, let us generically denote one of them by h. We compute an ℓ_1-normalized histogram h_i for each super-voxel r_i in the video. Two histograms h_i and h_j compared with histogram intersection, $s = \delta_1(h_i, h_j)$. The histograms are efficiently propagated through the hierarchy of super-voxels: Denoting $r_t = r_i \cup r_j$, the super-voxel obtained by merging the super-voxels r_i and r_j. We have

$$h_t = \frac{\Gamma(r_i) \times h_i + \Gamma(r_j) \times h_j}{\Gamma(r_i) + \Gamma(r_j)} \tag{1}$$

where $\Gamma(r)$ denotes the number of pixels in super-voxel r. The size of the new super-voxel r_t is $\Gamma(r_t) = \Gamma(r_i) + \Gamma(r_j)$.

Merging criterions based on size and fill: s_Γ, s_{F}. The similarity $s_\Gamma(r_i, r_j)$ aims at merging smaller super-voxels first:

$$s_\Gamma(r_i, r_j) = 1 - \frac{\Gamma(r_i) + \Gamma(r_j)}{\Gamma(\text{video})} \tag{2}$$

where $\Gamma(\text{video})$ is the size of the video (in pixels). This tends to produce super-voxels and therefore tubelets of varying sizes in all parts of the video (recall that we only merge contiguous super-voxels).

The last merging criterion s_{F} measures how well super-voxels r_i and r_j fit into each other. We define $B_{i,j}$ to be the tight bounding cuboid enveloping r_i and r_j. The similarity is given by

$$s_{\mathrm{F}}(r_i, r_j) = \frac{\Gamma(r_i) + \Gamma(r_j)}{\Gamma(B_{i,j})}. \tag{3}$$

Merging strategies. The merging strategy can be any of the individual merging criteria or it can be a sum of two or more of them. For instance, merging can be done based on only color similarity (s_{C}) or on only motion similarity (s_{M}); alternatively it can be done using a sum of color, motion and fill similarities ($s_{\mathrm{C}} + s_{\mathrm{M}} + s_{\mathrm{F}}$). Each merging strategy has a corresponding hierarchy, starting from n super-voxels, it leads to a new set of $n - 1$ super-voxels.

4. Motion features

Since we are concerned with action localization, we need to aggregate super-voxels corresponding to the action of interest, *i.e.*, points that deviate from the background motion due to camera motion. We can assume that usually later is the dominant motion in the image frame. The dominant (or global) image motion can be represented by a $2D$ parametric motion model. Typically, an affine motion model of parameters $\theta = (a_i)$, $i = 1...6$, or a quadratic model with 8 parameters can be used, depending on the type of camera motion and on the scene layout likely to occur:

$$\begin{aligned} w_\theta(p) =& (a_1 + a_2x + a_3y, a_4 + a_5x + a_6y) \\ \text{or } w_\theta(p) =& (a_1 + a_2x + a_3y + a_7x^2 + a_8xy, \\ & a_4 + a_5x + a_6y + a_7xy + a_8y^2), \end{aligned}$$

where $w_\theta(p)$ is the velocity vector supplied by the motion model at point $p = (x, y)$ in the image domain Ω. In this paper, we use the affine motion model for all the experiments.

4.1. Evidence of independent motion

First, we formulate the evidence that a point $p \in \Omega$ undergoes an independent motion at time step t. Let us introduce the displaced frame difference at point p and at time step t for the motion model of parameter θ : $r_\theta(p, t) = I(p + w_\theta(p), t + 1) - I(p, t)$. To simplify notation, we drop t when there is no ambiguity. At every time step t, the global parametric motion model can be estimated with a robust penalty function as

$$\hat{\theta} = \arg\min_\theta \sum_{p\in\Omega} \rho(r_\theta(p, t)), \tag{4}$$

where $\rho(.)$ is defined as the robust Tukey function [12]. To solve (4), we use the publicly available software Motion2D [18].

The robust function $\rho(r_\theta)$ produces a maximum likelihood type estimate: the so-called M-estimate [12]. Indeed, if we write $\rho(r_\theta) = -\log f(r_\theta)$ for a given function f, $\rho(r_\theta)$ supplies the usual maximum likelihood (ML) estimate. Since we are looking for independently moving objects in the image, we want to measure *the deviation* to the conformity with respect to the global motion. This is in spirit of the Fisher vector [19], where the deviation of local descriptors from a background GMM model is encoded to produce an image representation.

Figure 4. The original frame, its IME map and the result after segmentation are shown from left to right.

Let us consider the derivative of the robust function $\rho(.)$. It is usually denoted as $\psi(.)$ and corresponds to the influence function [12]. More precisely, the ratio $\psi(r_\theta)/r_\theta$ accounts for the influence of the residual r_θ in the robust estimation of the model parameters. The higher the influence, the more likely the point contributes to the global motion. Conversely, the lower the influence, the less likely the point contributes to the global motion. This leads us to define the *independent motion evidence* (IME) as

$$\xi(p, t) = 1 - \varpi(p), \tag{5}$$

where $\varpi(p)$ is the ratio $\frac{\psi(r_{\hat{\theta}}(p,t))}{r_{\hat{\theta}}(p,t)}$ normalized within $[0, 1]$.

4.2. Motion for segmentation

Each frame can be represented by the IMEs at each pixel, $\xi(p, t)$. The obtained IME frames are post-processed by applying morphological operations to obtain binary images. These binary images are applied as masks on the corresponding IME frames to obtain denoised IME maps. Applying GB video segmentation on sequences of these denoised maps partitions the video into super-voxels with independent motion. Therefore, we use it as an alternative for producing our super-voxels (Step 1 in Section 3). Figure 4 shows an example frame, its IME map and the result obtained by applying GB on the IME map. Thus resulting tubelets are more adapted to the action sequences, as evaluated in Section 5.1. This alternative for initial segmentation is also more efficient, about 4 times faster than GB on original video and produces 8 times fewer super-voxels.

4.3. Motion feature as merging criteria

We define a super-voxel representation for IME maps capturing the relevant information with efficiency. This representation is the histogram h_{M} involved in the merging criterion s_{M} mentioned in Section 3. We consider the binarized version of IME maps, *i.e.*, the binary images that results from morphological operations. At every point p, we evaluate the number of points q in its 3D neighborhood that are set to one. In a subvolume of $5 \times 5 \times 3$ pixels, this count value ranges from 0 to 75. The motion histogram $h_{\mathrm{M}i}$ of these values is computed over the super-voxel r_i. Intuitively, this histogram captures both the density and the compactness of a given region with respect to the number of points belonging to independently moving objects.

Merging Strategy	Video Segmentation			IME Segmentation		
	MABO	MR	#T	MABO	MR	#T
Initial voxels	36.2	0.4	862	48.6	28.0	118
M (s_{M})	56.2	43.2	299	52.9	35.7	90
C (s_{C})	47.3	24.3	483	51.1	35.1	93
T (s_{T})	44.6	23.4	381	51.2	38.8	81
S (s_{Γ})	47.8	23.5	918	52.2	35.2	158
F (s_{F})	50.9	30.7	908	52.7	38.8	155
M+S+F	57.2	49.8	719	54.2	40.3	129
T+S+F	52.6	34.0	770	53.9	46.3	145
C+T+S+F	53.4	38.4	672	54.5	45.2	127
M+C+T+S+F	58.1	48.6	656	55.1	41.5	122
Strategy set I	61.5	58.2	2346	56.6	48.3	469
Strategy set II	62.0	58.9	3253	56.8	49.5	625

Table 1. Mean Average Best Overlap for tubelet hypotheses using variety of segmentation strategies from UCF-Sports train set. [M:Independent motion evidence, C: Color, T: Texture, S: Size, F: 3D Fill, Strategy set I: {M, M+S+F, C+T+S+F, M+C+T+S+F}, Strategy set II: {M, F, M+S+F, C+T+S+F, M+C+T+S+F}].

5. Experiments

We evaluate our approach on two benchmarks that have localization groundtruth and have been evaluated for localization [4, 16, 24] : UCF-Sports [22] and MSR-II [4]. The first dataset consists of sports broadcasts with realistic actions captured in dynamic and cluttered environments. MSR-II contains videos of actors performing actions (handwaving, handclapping and boxing) in complex environments. It is suitable for cross-dataset experiment. As a standard practice, we use the KTH dataset for training. We first evaluate the quality of tubelet hypotheses generated by our approach. Then, Section 5.2 details our localization pipeline and compares our results with the state of the art methods on the two datasets.

5.1. Evaluation of Tubelet Quality: MABO

To evaluate the quality of our tubelet hypotheses, we compute the upper bound on the localization accuracy, as previously done to evaluate the quality of object hypotheses [29], by the Mean Average Best Overlap (MABO) and maximum possible recall (MR). Extending these measures to videos requires measuring the overlap between two sequences of boxes instead of boxes.

Localization score. In a given video V of F frames comprising m instances of actions, the i^{th} groundtruth sequence of bounding boxes is given by $gt^i = (B_1^i, B_2^i, ...B_F^i)$. If there is no action of i^{th} instance in frame f, then $B_1^i = \emptyset$. From the tubelet hypotheses, the j^{th} tubelet formed by a sequence of bounding boxes is denoted as, $dt^j = (D_1^j, D_2^j, ...D_F^j)$. Let $OV_{i,j}(f)$ be the overlap between the two bounding boxes in frame, f, which is computed as "intersection-over-union". The localization score between groundtruth tubelet gt^i and a tubelet dt^j is given by:

$$S(gt^i, dt^j) = \frac{1}{|\Gamma|} \sum_{f \in \Gamma} OV_{i,j}(f), \tag{6}$$

where Γ is the set of frames where at least one of B_f^i, D_f^j is not empty. This criterion generalizes the one proposed by Lan *et al*. [16] by taking into account the temporal axis. An instance is considered as localized or detected if the action is correctly predicted by the classifier and also the localization score is enough, *i.e*., $S(gt^i, dt^j) > \sigma$, the threshold for localization score.

The Average Best Overlap (ABO) for a given class c is obtained by computing, for each groundtruth annotation $gt^i \in G^c$, the best localization from the set of tubelet hypotheses $T = \{dt^j | j = 1 \dots m\}$:

$$\text{ABO} = \frac{1}{|G^c|} \sum_{gt^i \in G^c} \max_{dt^j \in T} S(gt^i, dt^j). \tag{7}$$

The mean ABO (MABO) synthesizes the performance over all the classes. Note that adding more hypotheses necessarily increases this score, so must be considered jointly with the number of hypotheses. Another measure for quality of localization used for images is maximum possible recall (*MR*). It is an upper bound on the recall with the given tubelet hypotheses. We also compare merging strategies using MR with a stringent localization threshold, $\sigma = 0.6$.

Table 1 reports the MABO, MR and the average number of tubelets (#T) for the train-set of the UCF-Sports dataset. Different strategies are compared for the two methods considered for initial segmentation (regular GB, and GB on IME). With regular GB segmentation, the best hypotheses are clearly produced by the strategies that include our s_{M} merging criterion: they attain the highest MABO and MR with the smallest number of tubelets. Many combinations of strategies were tried and the two best sets of strategies were chosen (described in Table 1). For the first chosen set, we achieve MABO=61.5% and MR=58.2% with only 2346 tubelets per video. Considering that the localization score threshold (σ) used in literature is 0.2, these MABO values are very promising.

The GB segmentation applied on our IME de-noised maps (See Section 4) generates a very good initial set (MABO = 48.6%). The MABO and specially MR further improve for all the strategies. Although the best values obtained, MABO=56.8% and MR=49.5%, are lower than those for the original video segmentation, the number of tubelets is only 625 on average. This is very useful for large videos where the number of samples, by sliding-subvolume or even by segmentation, is substantially higher.

For regular GB segmentation, MABO and MR are similar for both sets, so we choose strategy set I, as it needs lesser number of tubelets. With segmentation of IME maps, we choose strategy set II for its higher MR.

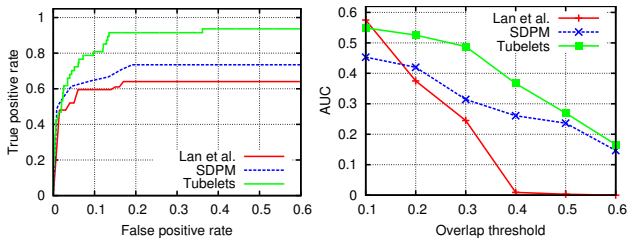

Figure 5. Comparison with concurrent methods [16, 24] on UCF-Sports: ROC at σ=0.2 and AUC for σ from 0.1 to 0.6.

5.2. Action localization

We now evaluate our tubelet hypotheses for action localization. With a relatively small number of candidate locations, our approach enables the use of expensive and powerful Bag-of-words based representation with large vocabulary sizes. We first extract state-of-art MBH descriptor computed along ω-trajectories using ω-flow [13][1]. We prefer using ω-trajectories over trajectories from optical flow [32] because they are more active on the actors, and also fewer trajectories are produced with ω-flow. To represent a tubelet, we aggregate all the visual words corresponding to the trajectories that pass through it. For training, we use a one-vs-rest SVM classifier with Hellinger (square-rooting+linear) kernel.

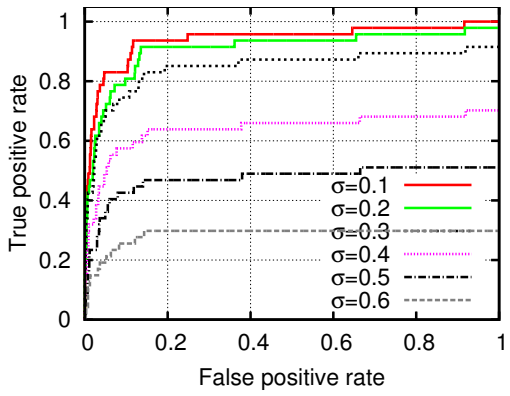

Figure 6. ROCs for σ from 0.1 to 0.6.

Experiments on UCF-Sports. This dataset consists of 150 videos with actions extracted from sports broadcasts. Ten action categories are represented, for instance "diving", "swinging-bench", "horse-riding", etc. We use the disjoint train-test split suggested by Lan *et al.* in [16]. The ground truth is provided as sequences of bounding boxes enclosing the actors. For training, we use the groundtruth tubelets and the tubelets provided by our method that have localization score greater than 0.7 with the groundtruth. Negative samples are randomly selected by considering tubelets whose overlap with ground truth is less than 0.2. We set the vocabulary size to $K = 500$ for Bag-of-words and use a spatial pyramid (1x1+2x2). We use the initial voxels from the GB segmentation performed on the original videos.

[1]Source code: `www.irisa.fr/texmex/people/jain/w-Flow`

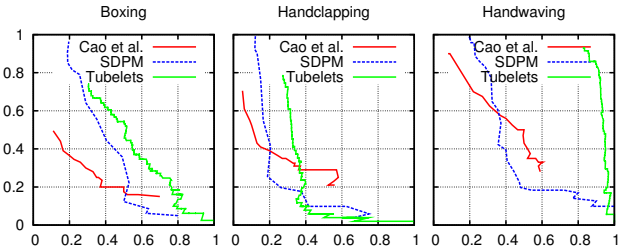

Figure 7. Precision/recall: Comparison with [4, 24] for the 3 classes on MSR-II. x-axis: precision, y-axis: recall.

For evaluating the quality of action localization, we follow the criteria explained in [16] and described in Section 5.1. Following previous works, we compare using the ROC curves and its AUC in Figure 5. On the left, we plot the ROC curve with $\sigma = 0.2$. In order to be consistent with SDPM and Lan *et al.*, we stop at FPR=0.6 and compute the AUC only for this part. On the right, we report AUCs for thresholds ranging from 0.1 to 0.6.

As can be seen from these figures, our approach significantly outperforms both methods. Figure 6 shows the complete ROC curves with different thresholds. We have almost total recall for $\sigma \leq 0.2$ and even for $\sigma = 0.5$ our recall is around 50%. Although our focus is localization, for classification we simply assign the video to the class to which the tubelet with maximum score is assigned. We obtain 80.24% of accuracy with this maximum score strategy. This is better than 79.4% of [22] and can be improved by specifically considering the classification task.

Experiments on MSR-II. This dataset consists of 54 videos recorded in a crowded environment, with many people moving in the background. Each video may contain one or more of three types of actions: boxing, handclapping and handwaving. An actor appears, performs one of these actions, and walks away. A single video has multiple actions (5-10) of different types, making the temporal localization challenging. Bounding subvolumes or cuboids are provided in the ground-truth. Since the actors do not change their location, it is as good as a sequence of bounding boxes. The localization criterion is subvolume-based, so we follow Cao *et al.* [4] and use the tight subvolume or cuboid enveloping tubelet. Precision-recall curves and average precision (AP) is used for evaluation [4]. Since MSR-II videos are much larger than UCF-Sports videos, to keep the number of tubelets low, we use the initial super-voxels from the GB segmentation of the IME maps along with strategy set II.

This dataset is designed for cross-dataset evaluation. Following standard practice, we train on the KTH dataset and test on MSR-II. While training for one class, the videos from other the two classes are used as the negative set. We compare with Cao *et al.* [4] and SDPM [24] in Figure 7.

Method	Boxing	Handclapping	Handwaving
Cao *et al*.	17.5	13.2	26.7
SDPM	38.9	23.9	44.7
Tubelets	**46.0**	**31.4**	**85.8**

Table 2. Average precisions for MSR-II

Table 2 shows that our tubelets significantly outperform the two other methods for all three classes.

6. Conclusions

We show, for the first time, the effectiveness of selective sampling for action localization in videos. Such hierarchical sampling produces category-independent proposals for action localization and implicitly covers variable aspect ratios and temporal lengths. Our independent motion evidence (IME) based representation of video provides a more efficient alternative for segmentation. The IME motion feature expresses both the individual density and the compactness of the action-related moving points in the super-voxel. An analysis shows that the proposed tubelet sampling heavily benefits from our motion features.

Overall, our approach outperforms the state of the art for action localization on two public benchmarks. As our method considers a significantly smaller number of candidate volumes at test time, we believe that our method will enable the use of more effective but also more costly representations of spatio-temporal volumes in future works.

Acknowledgments

We thank Inria international relationship and CS' Netherlands Prize for ICT Research for supporting the visit of Mihir Jain at the University of Amsterdam.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010.

[2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. IEEE *T-PAMI*, 2012.

[3] W. Brendel and S. Todorovic. Video object segmentation by tracking regions. In *ICCV*, 2009.

[4] L. Cao, Z. Liu, and T. S. Huang. Cross-dataset action detection. In *CVPR*, Jun. 2010.

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, May 2006.

[6] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012.

[7] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, 2010.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, Jun. 2010.

[9] I. Everts, J. van Gemert, and T. Gevers. Evaluation of color stips for human action recognition. In *CVPR*, 2013.

[10] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. IEEE *T-PAMI*, 32(9):1627–1645, 2010.

[11] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *ICCV*, 2011.

[12] P. Huber. *Robust statistics*. Wiley, New York, 1981.

[13] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, Jun. 2013.

[14] A. Kläser, M. Marszałek, C. Schmid, and A. Zisserman. Human focused action localization in video. In *Trends and Topics in Computer Vision*, pages 219–233, 2012.

[15] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, Jun. 2008.

[16] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, Nov. 2011.

[17] S. Manen, M. Guillaumin, and L. Van Gool. Prime Object Proposals with Randomized Prim's Algorithm. In *ICCV*, 2013.

[18] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Jal of Vis. Comm. and Image Representation*, 6(4):348–365, Dec. 1995.

[19] F. Perronnin and C. R. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[20] E. Rahtu, J. Kannala, and M. Blaschko. Learning a category independent object detection cascade. In *ICCV*, 2011.

[21] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, Jun. 2012.

[22] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, Jun. 2008.

[23] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, Jun. 2012.

[24] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, Jun. 2013.

[25] D. Tran and J. Yuan. Optimal spatio-temporal path discovery for video event detection. In *CVPR*, Jun. 2011.

[26] D. Tran and J. Yuan. Max-margin structured output regression for spatio-temporal action localization. In *NIPS*, Dec. 2012.

[27] D. Tran, J. Yuan, and D. Forsyth. Video event detection: From subvolume localization to spatio-temporal path search. IEEE *T-PAMI*, 2013.

[28] R. Trichet and R. Nevatia. Video segmentation with spatio-temporal tubes. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2013.

[29] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013.

[30] J. C. Van Gemert, C. J. Veenman, and J.-M. Geusebroek. Episode-constrained cross-validation in video concept retrieval. *IEEE Transactions on Multimedia*, 11(4):780–786, 2009.

[31] P. A. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[32] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, Jun. 2011.

[33] T. Wang, S. Wang, and D. Xiaoqing. Detecting human action as the spatio-temporal tube of maximum mutual information. IEEE *T-CSVT*, 24(2):277–290, 2014.

[34] C. Xu and J. Corso. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012.

[35] C. Xu, C. Xiong, and J. Corso. Streaming hierarchical video segmentation. In *ECCV*, 2012.

[36] J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. IEEE *T-PAMI*, 33(9):1728–1743, 2011.