Action Detection Survey

Zhijun Zhang

01/15/2020

Klaser et al. 2010 ECCV

Problems:

- Localize human actions both in space (the 2D image region) and time (the temporal window) in videos.

Challenges:

- Change of viewpoint, scale, lighting, partial occlusion, complex background, etc.

Introduce a new dataset, called 'Hollywood-localization'.[1]

[1] Alexander Kläser et al. "Human focused action localization in video". In: *ECCV*. Springer. 2010, pp. 219–233.

Klaser et al. 2010 ECCV

Motivation:

- Handle **more challenging datasets**, not only with static camera.
- Combine **human tracking** and classification for action localization.

Advantages:

- Improve the localization performance
- Reduce search complexity
- More efficient to learn new actions, since the tracks are agnostic.

Klaser et al. 2010 ECCV

Approach:

- Use **tracking-by-detection approach** to detect and track upper body of human actors.
  - HOG+SVM for **detection**;
  - KLT for **tracking**;
  - **interpolation** to produce smooth results;
  - **classification** to remove false positives.
- **Temporal classification** – Search a range of frames which contains the action.
  - **HOG-Track action descriptor**, extend 2D HOG to 3D;
  - **A two-stage SVM** classifier with sliding window.

Klaser et al. 2010 ECCV

	Drinking action	Smoking action
Drinking detector	54.1%	5.3%
Smoking detector	5.0%	24.5%

Table 1. Performance (AP) of drinking and smoking classifiers when localizing drinking and smoking actions. Note that the classifiers do not confuse the actions.

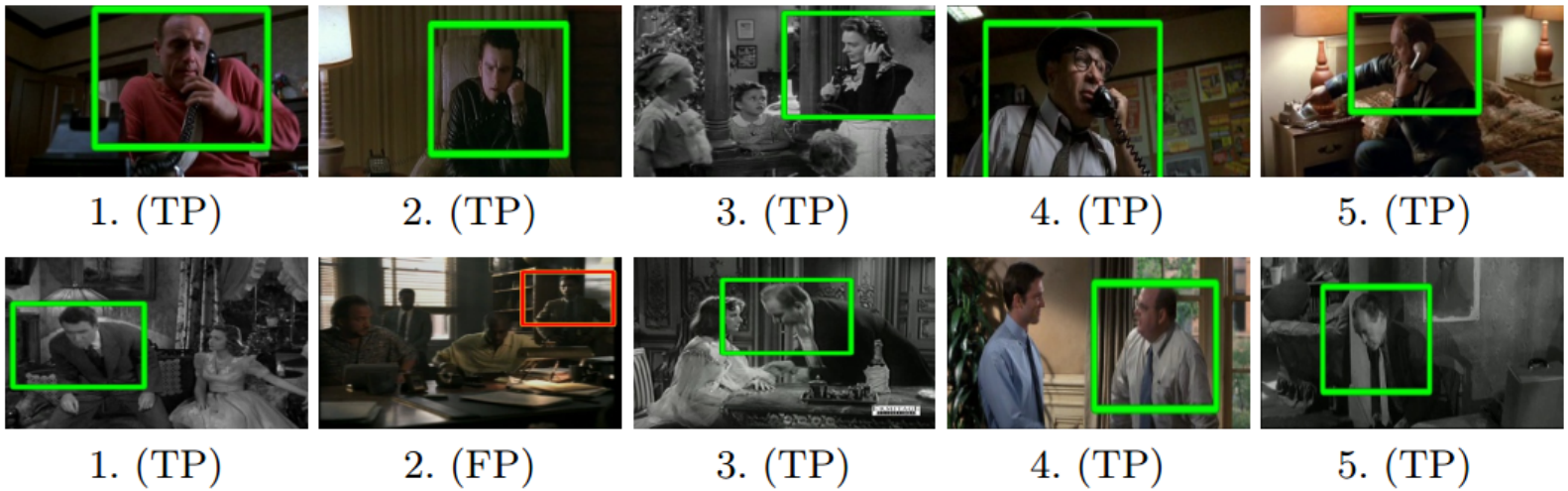


Fig. 9. The five highest ranked phoning (top) and standing-up (bottom) actions detected on *Hollywood–Localization*. For phoning the first FP is ranked 6th.

Klaser et al. 2010 ECCV

Comments:

- A simple pipeline to split the action localization into two parts: spatial localization and action temporal classification. Reduce the complexity of problems.
- Employ tracking method to localize human in videos.
- Not to distinguish each part of videos are beneficial for the action detection.
- Is the upper body of human reasonable? Sometimes actions happen in small areas, not always upper body.
- The dataset is quite specific and rather simple.

Laptev et al. 2007 ICCV

Problems:

- Recognition and localization of human actions in realistic scenarios.

Challenges:

- Individual variation of people in expression, posture, motion and clothing.
- Perspective effects and camera motions.
- Illumination variations.
- Occlusion and disocclusion
- The distracting effect of the scene surroundings.

Introduce a new dataset, called 'coffee cigarette' and drinking.[2]

[2] Ivan Laptev and Patrick Pérez. "Retrieving actions in movies". In: *ICCV*. IEEE. 2007, pp. 1–8.

Laptev et al. 2007 ICCV

Motivation

- Address action recognition and localization in movies with realistic variations of actions.
- For recognition, **access** the difficulty of the problem and the **relative performance of different methods**, such as shape-motion features, keyframe classifier, key points, etc.
- For detection, **combine keyframe classifier and space-time classifier** to detect action.

Laptev et al. 2007 ICCV

Investigate three classifiers.

- Space-time action classifier:
  - Use Adaboost with Fisher discriminant.
  - Motion and shape Features, mainly based on histogram of spatial gradient and optical flow.
- A keyframe classifier using boosted histogram.
- Space-time interest points(STIP) classifier using NN method.

Observation

- Approaches perform differently in various videos, which inspire the combination of complementary classifiers.

Laptev et al. 2007 ICCV

Keyframe priming for action detection

- Apply keyframe detector to all positions, scales, frames;
- Generate action hypotheses in terms of space-time block;
- Each block is classified according to the space-time classifier.

Laptev et al. 2007 ICCV

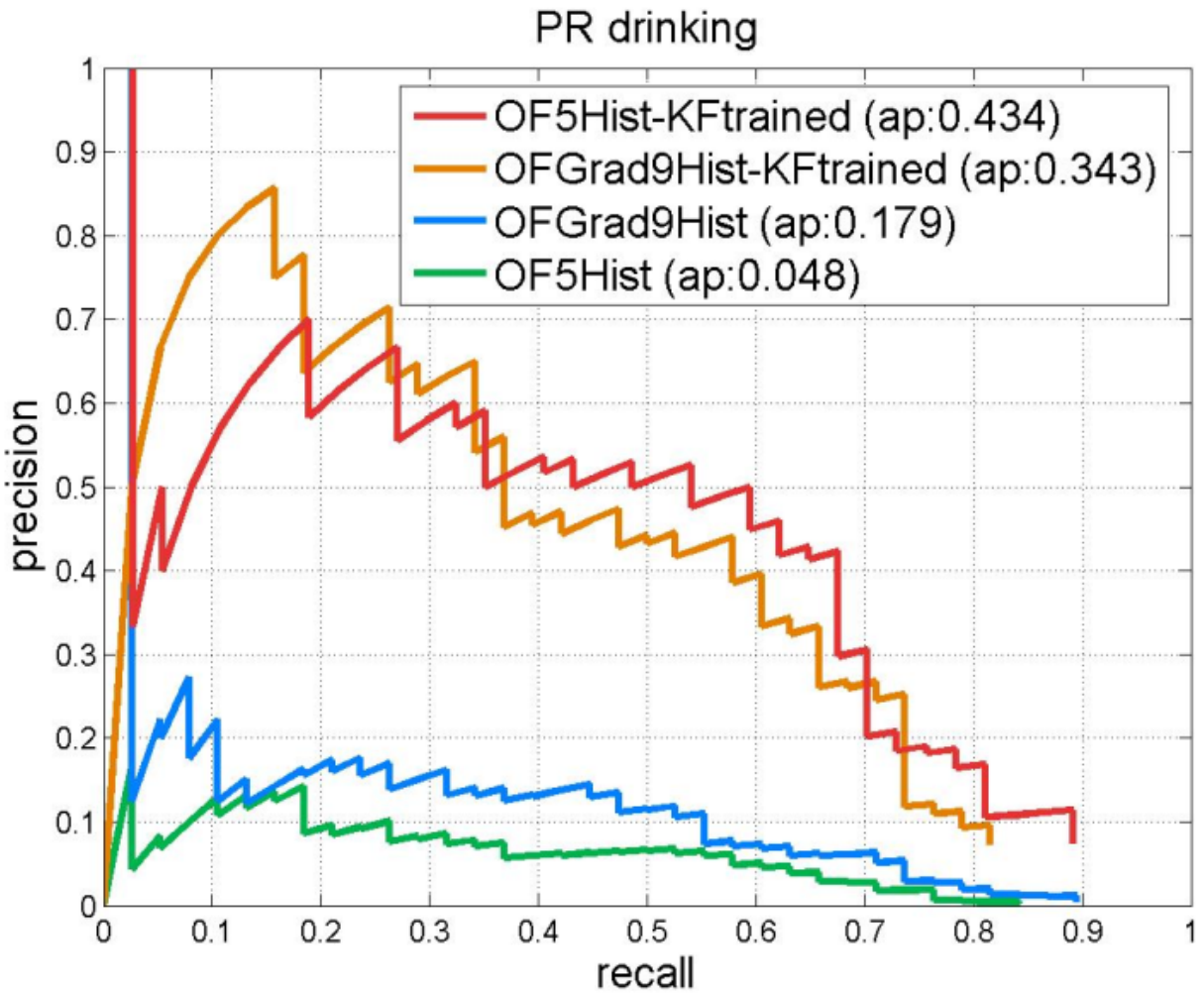


Figure 7. Precision-recall curves illustrating the performance of drinking action detection achieved by the four tested methods.

Laptev et al. 2007 ICCV

- A investigation of different approaches to action recognition.
- Localization and temporal recognition pipeline.
- Very specific actions.
- How to link the human to produce stable tube.
- What's the choice of different temporal-space features.

Ke et al. 2007 ICCV

Problems:

- Video event detection in crowded, dynamic environments, such as a person waving his or her hand.

Challenges:

- Occluded by other people or objects.
- Distracted motion, optical flow may be dominated by other distracting objects.
- Contain multiple moving objects and significant clutter.

34

[3] Yan Ke, Rahul Sukthankar, and Martial Hebert. "Event detection in crowded videos". In: *ICCV*. IEEE. 2007, pp. 1–8.

[4] Yan Ke, Rahul Sukthankar, and Martial Hebert. "Volumetric features for video event detection". In: *IJCV* 88.3 (2010), pp. 339–362.

Ke et al. 2007 ICCV

Past approaches

- Tracking-based approaches cannot produce stable results in crowded environments, may be noisy.
- Flow-based approaches is not stable.
- Shape-based rely on figure-ground segmentation, which is stable in crowded scenes.
- Space-time interest points fail to capture smooth motions and tendency to generate spurious detections at object boundaries.

Motivation

- Effective representation of shape and motion for event detection;
- Effective matching of event models to over-segmented spatio-temporal volumes, to avoid figure-ground separation.

Ke et al. 2007 ICCV

Approaches:

- **Shape matching** between the template and the video frame, which is **over-segmented**, to avoid unreliable foreground separation, and invariant to appearance and lighting changes.
- **Flow matching** complements the shape descriptor.
- **Matching parts** to make it more **robust** to the spatial and temporal variability.

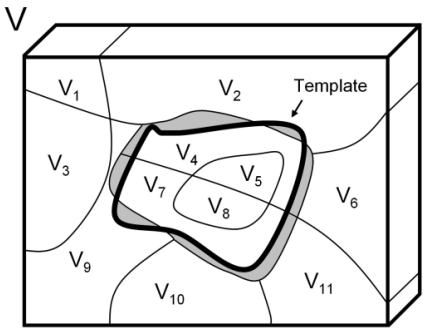


Figure 4: Example showing how a template is matched to an over-segmented volume using our shape matching method. The template is drawn in bold, and the distance (mismatch) is the area of the shaded region.

(a) Whole (b) Parts

Figure 5: To generalize the model and allow for more variability in the action, we break the action template (a) into parts (b). The model can be split in both space or time to generate the parts.

Ke et al. 2007 ICCV

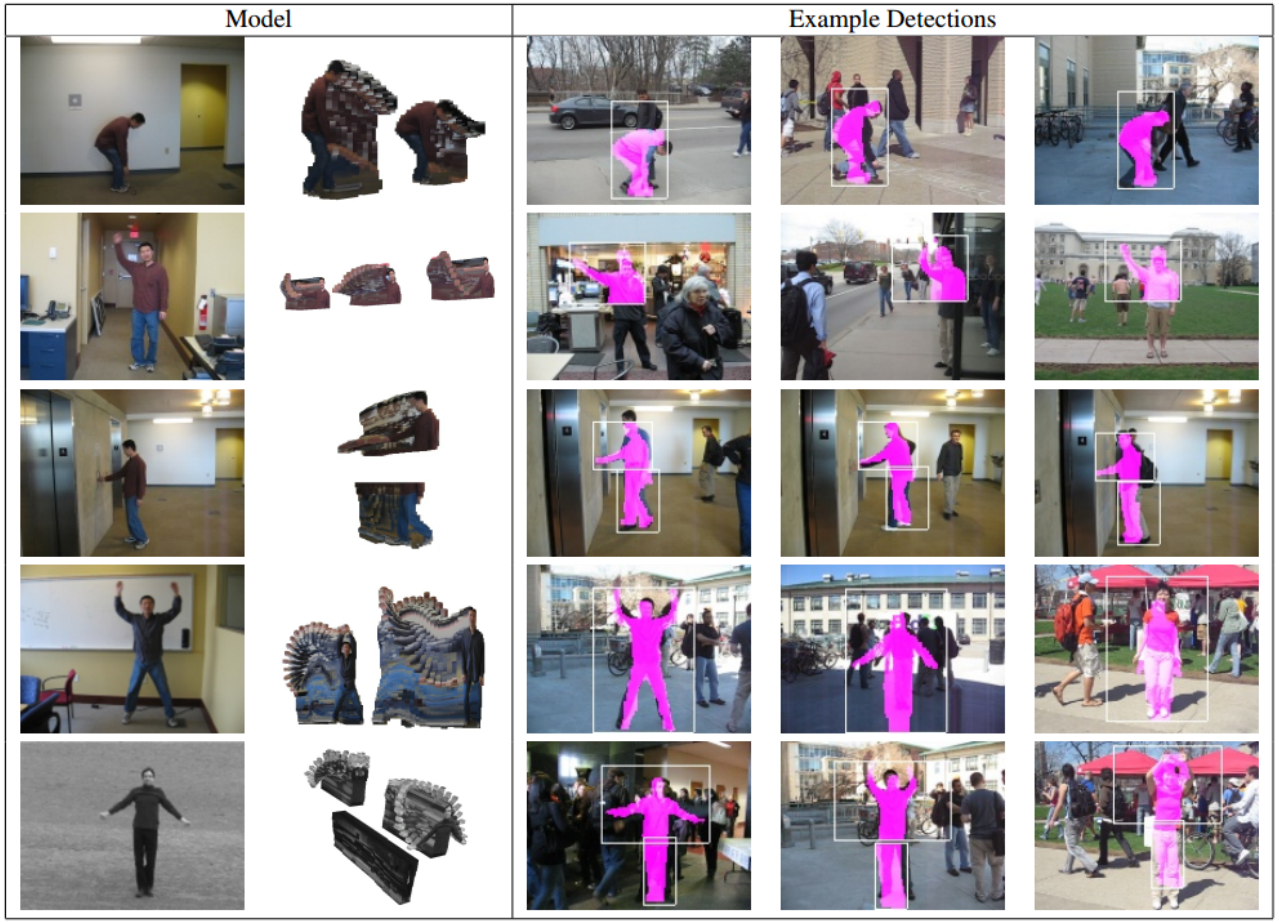


Figure 7: Examples of event detection in crowded video. Training sequences and event models are shown on the left. Detections in several challenging test sequences are shown on the right. The action mask from the appropriate time in the event model is overlaid on the test sequence frame, and a bounding box marks the matched location of each part.

Ke et al. 2007 ICCV

Comment:

- **Emphasizes the matching aspects of event detection** and demonstrates robust performance on real-world videos.
- They **assume that the shape of actions are similar**, so they can use shape-matching to recognize the action. **But actions may happen in various shapes and scales due to change of people, view point**, etc.
- Besides, **the start and end of the action may be quite different from the key frame of action**, how can we recognize the start and end of the action based on the shape of key frame.
- Only apply on the simple typical actions, which can be recognized by **the shape of human**.

Yuan et al. 2009 CVPR

Problems:

- Actions are **spatio-temporal patterns** which can be characterized by collection of spatio-temporal invariant features.
- Detection of actions is to find the **re-occurances (e.g. through pattern matching)** of such spatio-temporal patterns.

Challenges:

- Searching for actions in the video space is much more complicated than searching for objects in the image space, due to **enormous search space**(3D volume).
- Human actions involve **tremendous intra-pattern variations**, caused by performing speed, clothing, scale, view points, and partial occlusion. Single template may not handle the problem, but multiple templates increase the computational costs.

56

[5] Junsong Yuan, Zicheng Liu, and Ying Wu. "Discriminative subvolume search for efficient action detection". In: *CVPR*. IEEE. 2009, pp. 2442–2449.

[6] Junsong Yuan, Zicheng Liu, and Ying Wu. "Discriminative video pattern search for efficient action detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.9 (2011), pp. 1728–1743.

Yuan et al. 2009 CVPR

Motivation

- Formulate action detection as a problem of **searching the 3D subvolumn that has maximum total votes, i.e. maximum mutual information**, w.r.t the action class.

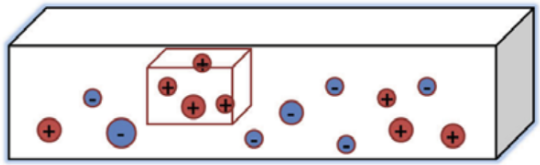


Figure 1. Action detection is formulated as searching for a subvolume in video that has the maximum mutual information toward the action class. Each circle represents a spatio-temporal feature point which contributes a vote based on its own mutual information.

Three benefits:

- Well handle action variations by **using all of the training data** instead of a single template.
- Pure data-driven approach **without requiring object tracking and detection**.
- 3D videos is computationally **efficient** and is suitable for a **real time system implementation**.

Yuan et al. 2009 CVPR

Formulation

- **Employ STIPs(extension of SIFT in 3D) of invariant features**, and use HOG and HOF to describe them.
- Assuming STIPs are **independent**, the action classfier can be easily determined by maximizing sum of mutual information of each component in the video clip. (**naive-Bayes based on mutual information maximization**)
- **Decouple the temporal and spatial spaces** and applies different search strategies on them to **speed up the search**.
- **Discriminate matching** can be regarded as the use of two template classes, positive data and negative samples.

Yuan et al. 2009 CVPR

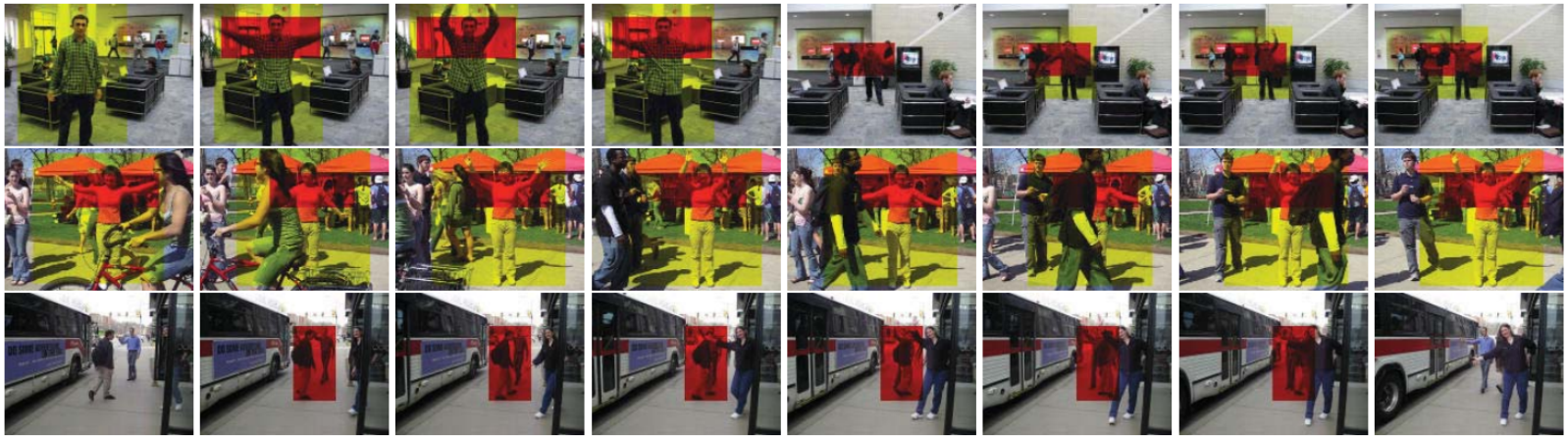


Figure 5. Detection results of two-hand waving. The yellow bounding box is the ground truth label of the whole human body action and the red bounding box is our detection of two-hand waving. The 1_{nd} row: detection under multiple spatial scales; the 2_{rd} row: detection with partial occlusions; the 3_{th} row: a false detection caused by two individual hand-wavings from two different persons. More results can be seen in the supplementary materials.

Yuan et al. 2009 CVPR

- A **pure data-driven approach**, which can utilize the multiple samples, instead of single templates.
- Based on STIPs features, **not require any tracking or detection methods**.
- **A new formulation**, which is quite reasonable, which can handle multiple variations..
- **Solid prove** for the upper bound.
- But **the assumption of independence of each STIP may not be reasonable** in real scenes. Can we model the relationship between the keypoint features.
- And they **only model local features**, is the other global or wider features can be utilized?

Ikizler et al. 2007 ICCV

Problems:

- Collection of motions using a simple and effective query language, without examples.

Difficulties:

- Good kinematic tracking is hard;
- Models typically have too many parameters to be learnt directly from data;
- For much everyday behaviour, there isn't a taxonomy.

78

[7] Nazli Ikizler and David Forsyth. "Searching video for complex activities with finite state models". In: *CVPR*. IEEE. 2007, pp. 1–8.

[8] Nazlı İkizler and David A Forsyth. "Searching for complex human activities with no visual examples". In: *International Journal of Computer Vision* 80.3 (2008), pp. 337–357.

Ikizler et al. 2007 ICCV

Motivation

- **Develop a vocabulary or develop expressive tools for authoring models**, which can **composite complex actions** and robust as possible to view effects and to details of appearance of the body.
- Distinguish motions information based on timescale: **short-timescale representation (acts)**; **medium timescale actions**, like walking, running, jumping, standing, waving, whose temporal extent can be short (but may be long) and are typically composites of multiple acts; and **long timescale activities**, which are complex composites of actions.

Ikizler et al. 2007 ICCV

Review on three threads to interpret activities.

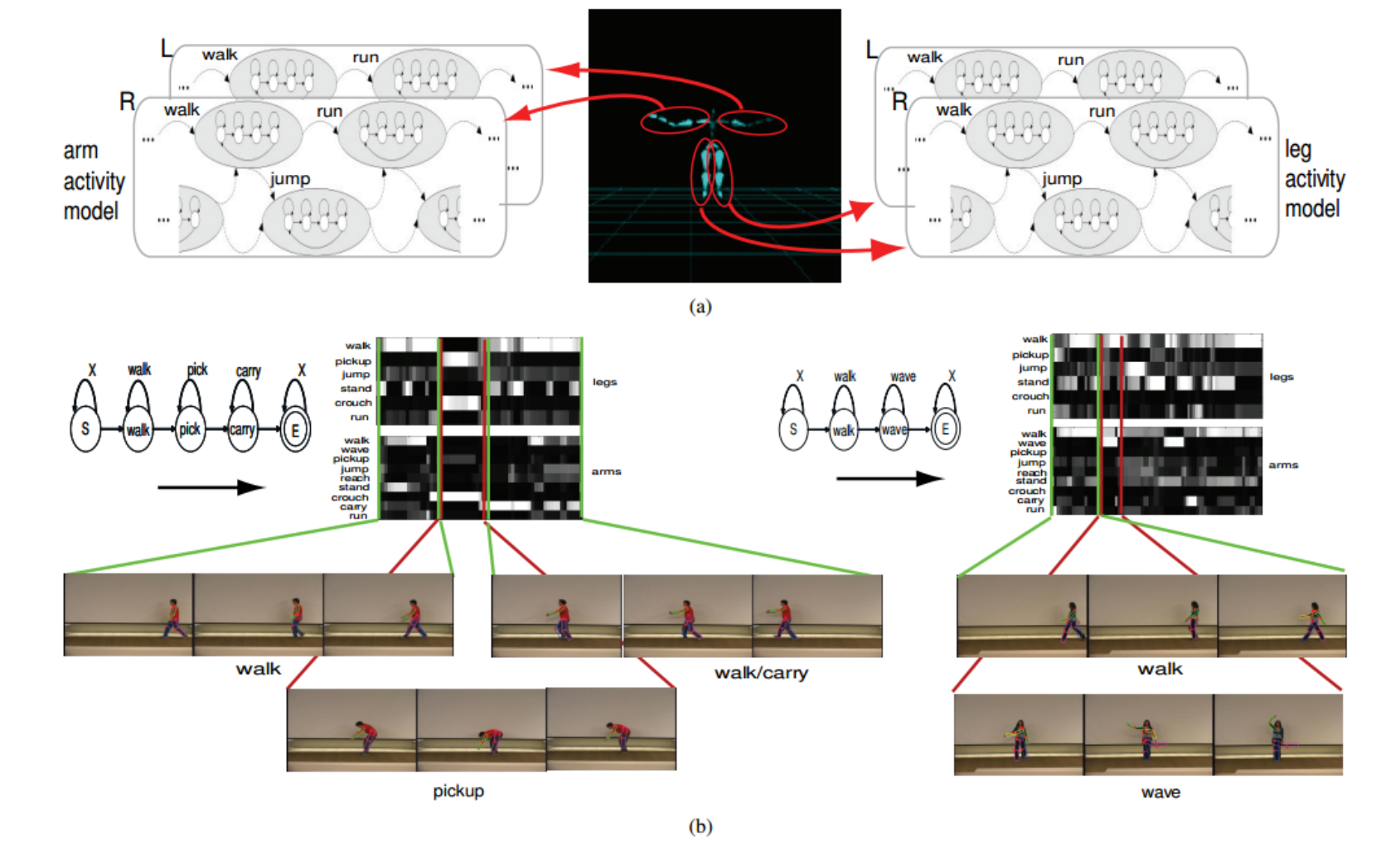
- Use **temporal logics** to represent crucial order relations between states that constrain activities;
- Use spatio-temporal **templates** to identify instances of activities;
- Use **hidden Markov models of dynamics**.

Ikizler et al. 2007 ICCV

Approach:

- Use tracking and snippet models to obtain 3D information of body (limbs).
- Use **generative model HMM** to represent the dynamics of motion on each limbs, to composite a action.
- Use a complex larger HMM to model the activity, so that many of actions or activities can be represented by a vocabulary, which is learned by observed data.

Ikizler et al. 2007 ICCV



Ikizler et al. 2007 ICCV

Comments:

- Model the configuration of human body and its evolvement in time domain.
- You **never know each action or activity in real world can be modeled as a fix taxonomy**.
- But **generative model** is quite novel for me to think about action recognition problems.
- If **there exist a feature space**, where every actions can be modeled as a fix order?
- **The speech theory can be transferred into this field.**
- Use **logic knowledge** to represent an action may be a high-level thought.

Jiang et al. 2006 CVPR

Problems:

- Finding actions of people in an uncontrolled video.

Difficulties:

- There is no effective way to segment an object in such videos.
- Highly articulated character of the human body.
- Large variability of clothing
- Strong background clutter.

9

[9] Hao Jiang, Mark S Drew, and Ze-Nian Li. "Successive convex matching for action detection". In: *CVPR*. vol. 2. IEEE. 2006, pp. 1646–1653.

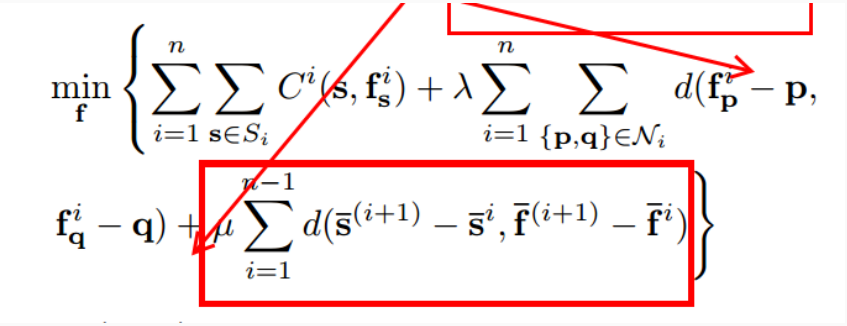
Jiang et al. 2006 CVPR

Motivation:

- Represent an action as a sequence of body postures with specific temporal constraints.
- Formulate matching problem as an energy minimization problem.
- Include the center continuity constraint, which is important to force the matching stick to one object.

Jiang et al. 2006 CVPR

Formulation



$$\min_{\mathbf{f}} \left\{ \sum_{i=1}^{n} \sum_{\mathbf{s} \in S_i} C^i(\mathbf{s}, \mathbf{f}_{\mathbf{s}}^i) + \lambda \sum_{i=1}^{n} \sum_{\{\mathbf{p},\mathbf{q}\} \in \mathcal{N}_i} d(\mathbf{f}_{\mathbf{p}}^i - \mathbf{p}, \mathbf{f}_{\mathbf{q}}^i - \mathbf{q}) + \mu \sum_{i=1}^{n-1} d(\bar{\mathbf{s}}^{(i+1)} - \bar{\mathbf{s}}^i, \bar{\mathbf{f}}^{(i+1)} - \bar{\mathbf{f}}^i) \right\}$$

Matching cost + intra-frame smooth constraints + inter-frame center continuity constraint.

- Choose edge transformed features, which is robust to color and deformation.
- Use basis target points to represent feature points, speeding up the algorithm.
- Long long math.....

Jiang et al. 2006 CVPR

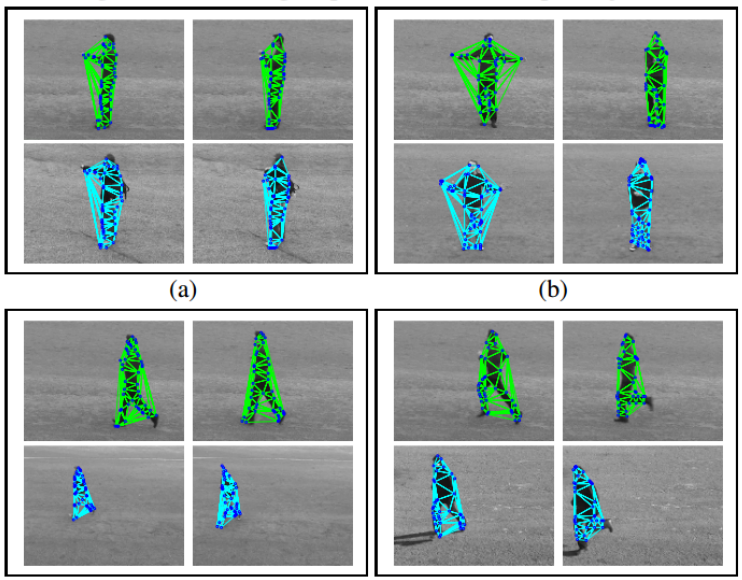


Figure 6. Matching examples. In (a, b, c, d) top rows are templates and bottom rows are matching results.

Figure 8. Searching gesture “work” in a 1000-frame sign language sequence. (a) Templates; (b..h) Top 7 matches of the shortlist.

Jiang et al. 2006 CVPR

Comments:

- Present a successive convex programming scheme to match video sequences using intra-frame and inter-frame constrained local features
- The matching scheme has unique features in searching: it involves a very small number of basis points and thus can be applied to problems that involve large number of target points.
- Energy minimization modeling for multi-feature matching with templates.
- Ignore the temporal relationship among the action.
- Ignore the temporal features.
- Just like an object matching with multiple templates.

Seo et al. 2009 ICCV

Problems (Unsupervised action detection):

- Given a single ”query” video of an action of interest (for instance a short ballet turn), and we are interested in detecting similar actions within other ”target” videos.

Challenges:

- Similar actions with different clothes, or in different illumination and background can result in a large appearance variation.
- The same actions performed by two different people may look dissimilar in terms of action speed or frame rate of the video.

10

[10] Hae Jong Seo and Peyman Milanfar. “Detection of human actions from a single example”. In: *ICCV*. IEEE. 2009, pp. 1965–1970.

Seo et al. 2009 ICCV

Motivation

- **Training-free** action analysis trend (provided only a single query video)
- Inspired by great progress of **adaptive kernel regression** for image denosing, interpolation, deblurring, and 2-D generic object detection.
- **Space-time local steering kernel** can capture the essential **local behavior of a spatio-temporal neighborhood**.

$$K(\mathbf{x}_s - \mathbf{x}) = \frac{\sqrt{\det(\mathbf{C}_s)}}{h^2} \exp\left\{\frac{(\mathbf{x}_s - \mathbf{x})^T \mathbf{C}_s (\mathbf{x}_s - \mathbf{x})}{-2h^2}\right\}$$

Seo et al. 2009 ICCV

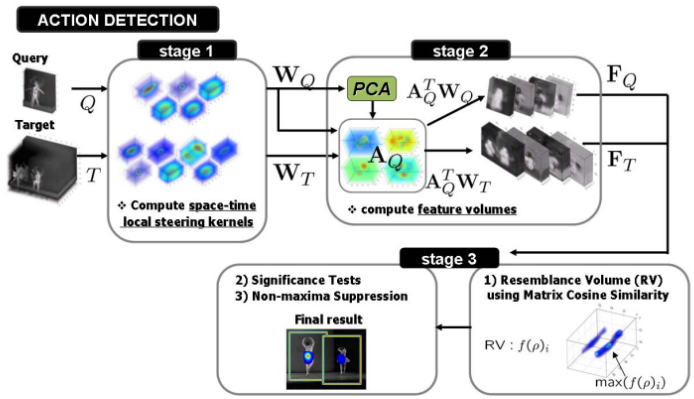


Figure 2. Action detection system overview (There are broadly three stages.)

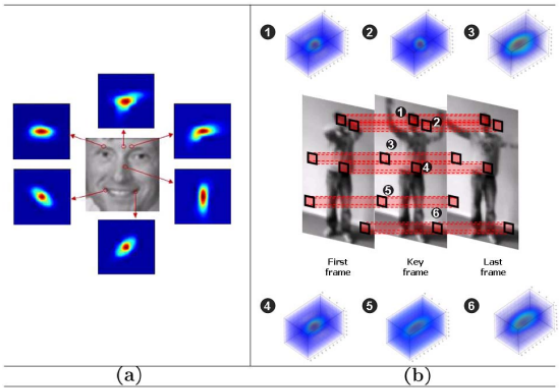


Figure 3. (a) Examples of 2-D LSK in various regions. (b) Examples of space-time local steering kernel (3-D LSK) in various regions. Note that key frame means the frame where the center of 3-D LSK is located.

Seo et al. 2009 ICCV

Experiments:

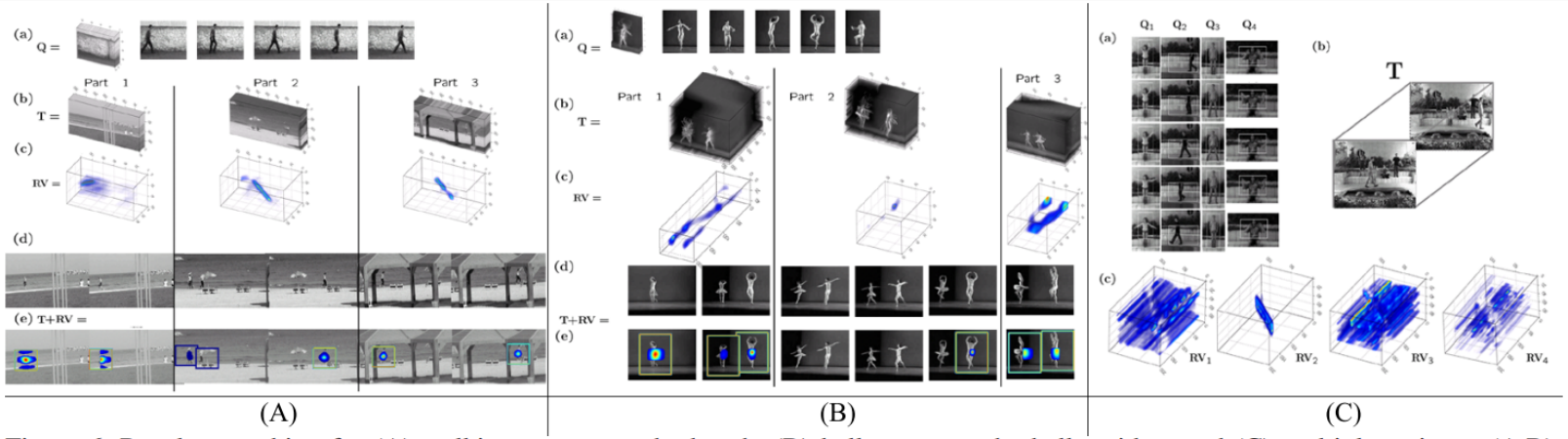


Figure 6. Results searching for (A) walking person on the beach, (B) ballet turn on the ballet video, and (C) multiple actions. (A,B): (a) query video (a short walk clip) (b) target video (c) resemblance volumes (RV) (d) a few frames from T (e) frames with resemblance volume on top of it. (C): (a) four different short video queries. Note that white boxes represent actual query regions (b) target video T (c) resemblance volumes (RV)s with respect to each query.

Seo et al. 2009 ICCV

Comments:

- A new features called space-time local steering kernel, a nonparametric method.
- Is video query searching reasonable?
- So many abundant information, how to know what exactly we want?

Hu et al. 2009 ICCV

Problems:

- The detection of semantic human actions in complex scenes, which suffers from cluttered background, heavy crowds, occluded bodies, and spatial-temporal boundary ambiguous.

Figure 1. Illustration of the action detection problem in complex scenes. The image in the first row shows a crowded scene of a shopping mall. The left five images in the second row are the detected actions of reaching (first two), pointing, squatting, and bending to merchandise on shelf. The last image in the second row is a negative sample as the customer is only walking in front of the shelf.

11

[11] Yuxiao Hu et al. "Action detection in complex scenes with spatial and temporal ambiguities". In: *ICCV*. IEEE. 2009, pp. 128–135.

Hu et al. 2009 ICCV

Motivation

- Address the problem of **spatial ambiguous** (hard to locate human body precisely) and **temporal ambiguous** (duration may vary within an action, hard to decide the start or end point of actions)

- Employ **multi-instance learning (MIL) based Support Vector Machine (SVM)** to handle these ambiguities in both spatial and temporal deomain.

Figure 2. Illustration of multi-instance learning for human action detection in complex scenes. Top: spatial ambiguities in scale and position. Bottom: temporal ambiguities in time domain. Different color boxes indicate different candidate instances. For better viewing, please see the original pdf file.

Hu et al. 2009 ICCV

Approach:

- Labelling the video frame.
- Generate multiple segments by cropping at various locations/scales and different time.
- Employ motion history image (MHI) feature, foreground image (FI) and HOG, to consider both motion and appearance features.
- Propose an algorithm named SMILE-SVM (Simulated annealing Multiple Instance LEarning Support Vector Machine), which aims to obtain a global optimum.

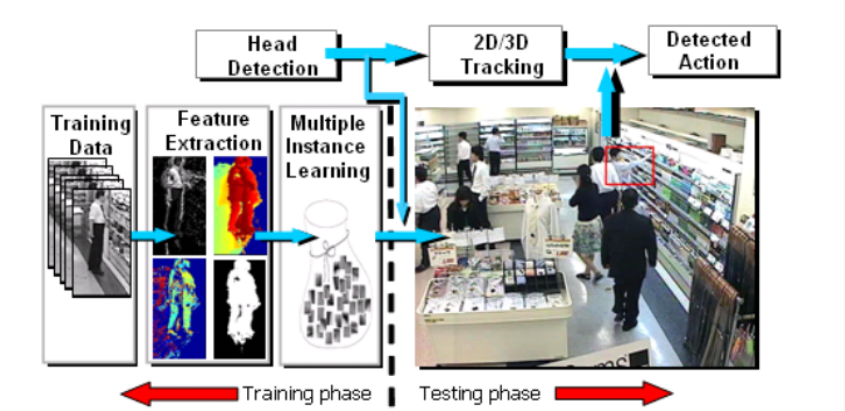


Figure 6. Overview of the system for action detection. Note that by combining our action detection result with a 3D tracker, the shop manager may know customers' interests over different shelves.

Hu et al. 2009 ICCV

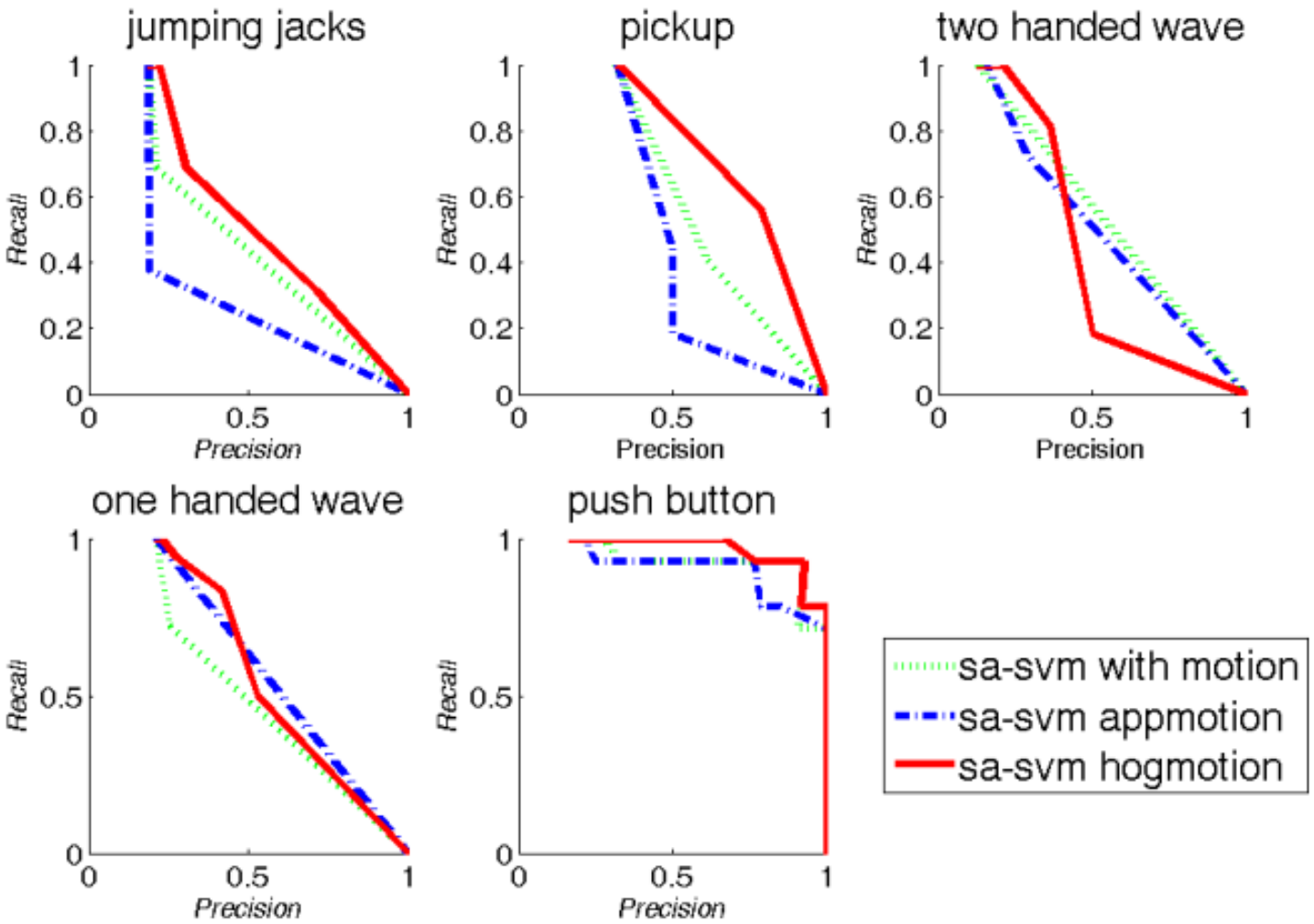


Figure 5. Comparison precision/Recall curves for a variety of actions and features. Mhi+HOG feature on SMILE-SVM outperforms other features in most cases.

Hu et al. 2009 ICCV

- A very straight-forward discriminative approach to handle spatio-temporal ambiguous with multiple candidates.
- But they treat each frame individual, not to discuss the temporal relationship between frames, they assume the action of each frame can be likely to the training data.
- Only handle the action. which is similar with the historic action. Maybe fail in some big variety actions.

Yu et al. 2011 CVPR

Problems:

- Given a database with N video clips. These video clips may contain various types of actions such as handwaving, boxing.
- Our objective is, given one or more query videos, to extract all the sub-volumes which are similar to the query.

Challenges:

- Usually **only a single query example** is provided, in contrast to many labelled data in action recognition and detection. Furthermore, exist possible action variations.
- A retrieval system must have **a fast response time** because otherwise the user experience would suffer.
- A retrieval process typically **involves user interactions**, which allows the user to clarify and update their preferences.

[12][13]

[12] Gang Yu, Junsong Yuan, and Zicheng Liu. "Unsupervised random forest indexing for fast action search". In: *CVPR*. IEEE. 2011, pp. 865–872.

[13] Gang Yu et al. "Fast action detection via discriminative random forest voting and top-k subvolume search". In: *IEEE Transactions on Multimedia* 13.3 (2011), pp. 507–517.

Yu et al. 2011 CVPR

Motivation

- Each video is characterized by a **collection of interest points**.
- Introduce **a random forest** to construct indexing for these interest points, which can **speed up**.
- Easily **extend** to support interactive search by **incrementally** adding user labeled actions to the query set.

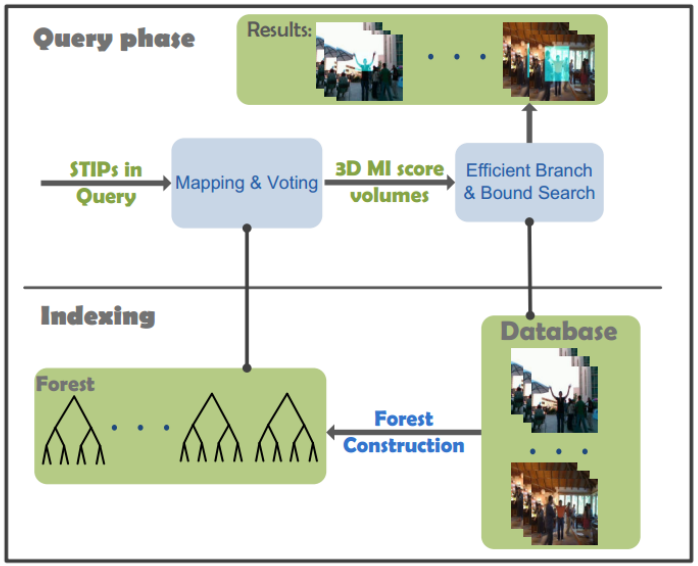


Figure 1. Overview of our algorithm.

Yu et al. 2011 CVPR

Appproaches

- HOG+HOF features
- Random forest based indexing.
- Coarse-to-fine subvolume search scheme.
- Refinement with Hough Voting
- Interactive search

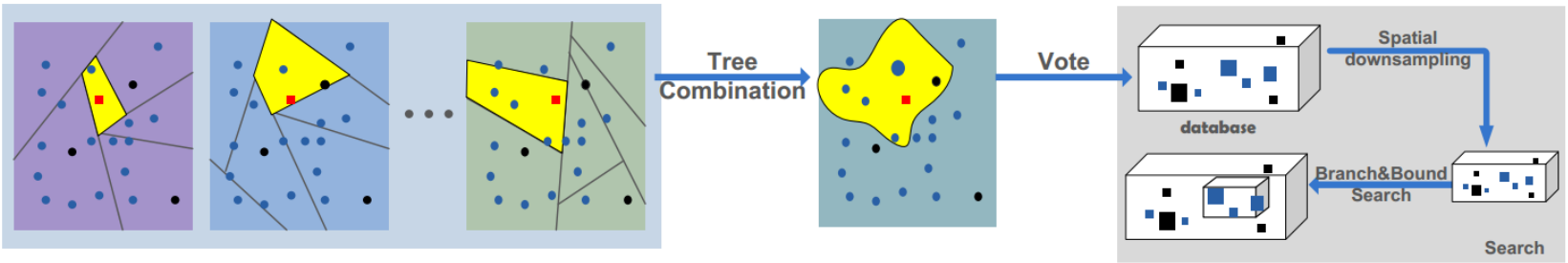


Figure 2. A schematic illustration of random forest based indexing and action search.

Yu et al. 2011 CVPR

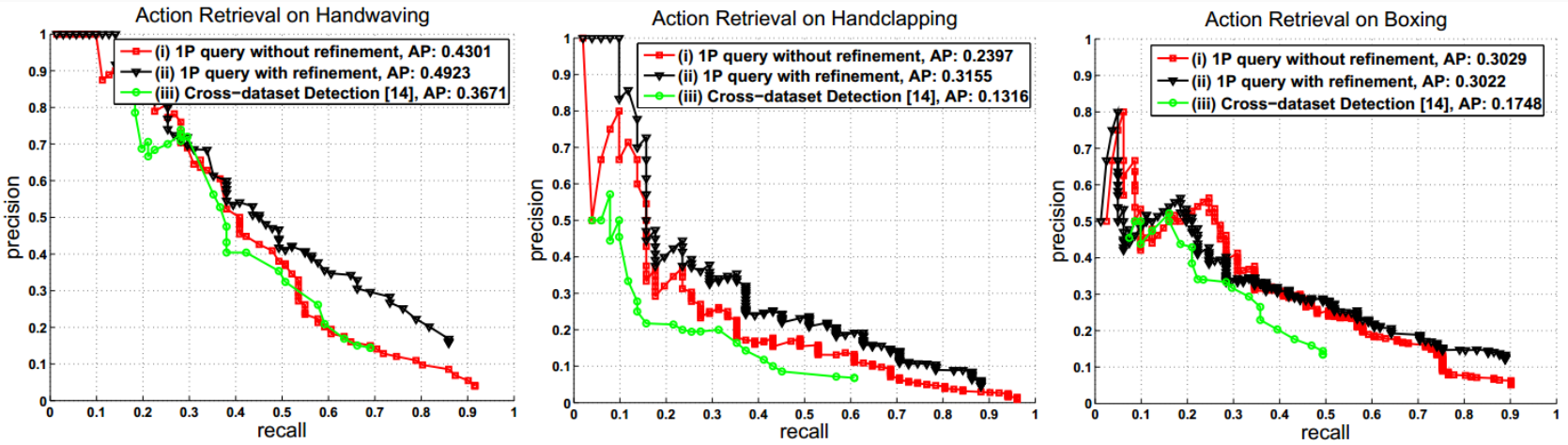


Figure 4. Precision-recall curves for action search.

Yu et al. 2011 CVPR

Comments:

- The paper focuses on the **search efficiency and accuracy**. Not to focus on the features and metrics.
- **Faster** and **easily extendable** by interactive action search.
- Extension of Yuan's work.

Cao et al. 2010 CVPR

Problems:

- Cross-dataset action detection, which aims to generalize action detection models built from a source dataset to a target dataset.
- Adapt the existing classifier from a source dataset to a new target dataset, while requiring only a small amount of labeling samples or even no labels at all.

Challenges:

- Change of performers, background, lighting condition, scales, and performing speeds.

14

[14] Liangliang Cao, Zicheng Liu, and Thomas S Huang. "Cross-dataset action detection". In: *CVPR*. IEEE. 2010, pp. 1998–2005.

Cao et al. 2010 CVPR

Motivation (two factors):

- Actions in different datasets still share similarities to some degree.
- The classifier adaptation can leverage the spatial and temporal coherence of the individual actions in a new dataset.

Main idea

- Since STIPs from different datasets may not have the same distribution, we employ **a probabilistic representation of the original STIPs**.
- Introduce **a prior distribution of the GMM parameters and propose an adaptation approach** to incorporate to prior information for cross-dataset analysis.

Cao et al. 2010 CVPR

Algorithm 1 Cross-dataset Action Detection

1: **Input**: labeled source dataset S and target dataset T.
2: Train background model $Pr(\theta_b)$ based on all the STIPs in the T.
3: In **Source** dataset S:
4: apply (3) to S and obtain θ_c.
5: In **Target** dataset T:
6: update **Q** using (4).
7: update θ_c using (3).
8: repeat the last two steps for several rounds.
9: **Output** the action model and the detected regions in T.

Cao et al. 2010 CVPR

Table 1. Comparing the accuracy on KTH

Work	Accuracy	Num of training
Schuldt *et al.* [17]	71.71%	16 persons
Dollar *et al.* [2]	80.66%	16 persons
Niebles and Fei-Fei [15]	83.92%	16 persons
Huang *et al.* [6]	91.6%	16 persons
Laptev *et al.* [10]	91.8%	16 persons
Yuan *et al.* [22]	93.3%	16 persons
Liu and Shah [12]	94.16%	16 persons
Our work	**95.02%**	16 persons
Our work	**94.01%**	8 persons
Our work	**90.63%**	4 persons

Figure 6. Action detection results on TRECVID subdataset. The running person is bounded by a red box. In the first row, one person was running in the crowd with a baby cart. In the second row, another person first ran and then walked away. Our approach detects the running action only.

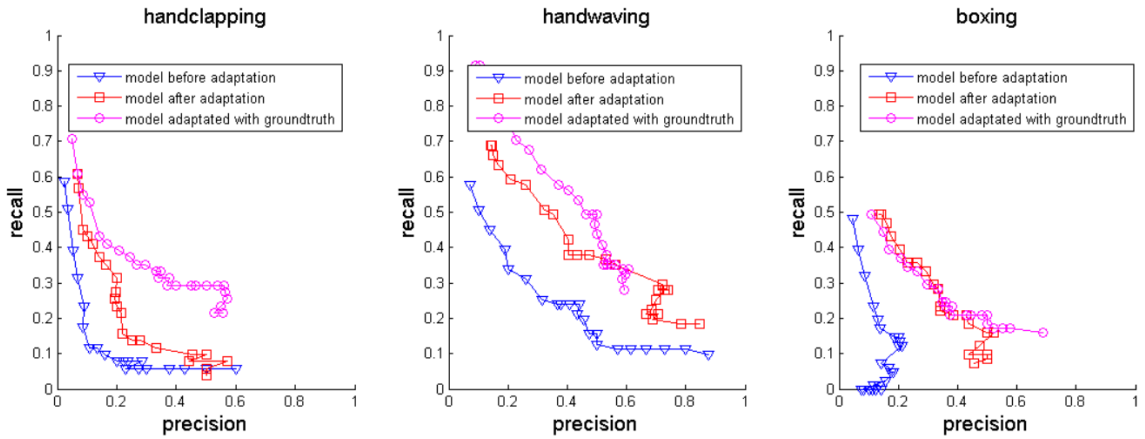


Figure 5. Adaptive action detection on MSR dataset. We compare two kinds of adaptations methods: Adaptation using cross-dataset and adaptation using a small amount of groundtruth labels. Both approaches outperform the original method significantly.

Cao et al. 2010 CVPR

- Seamlessly Incorporate the information from a source dataset to a target dataset, by introducing the prior of GMM parameters.
- **A meaningful topic** about cross-data action. Can we further explore the common property of different datasets or actions, such as configuration?
- The assumption that the probalistic of STIPs are similar, maybe failed in many various scenes.
- The paper inspired me to think of the similarity of same actions in different scenes or different datasets.

Tran et al. 2012 NIPS

Problem:

- Adapt **structured regression** in action localization.
- Structured regression makes great progress in object localization, which **considers the correlations** among the output variables and **avoids an exhaustive search** of the subwindows.

Challenges:

- Action could occur in a **exponentially large size** of the structured video space.

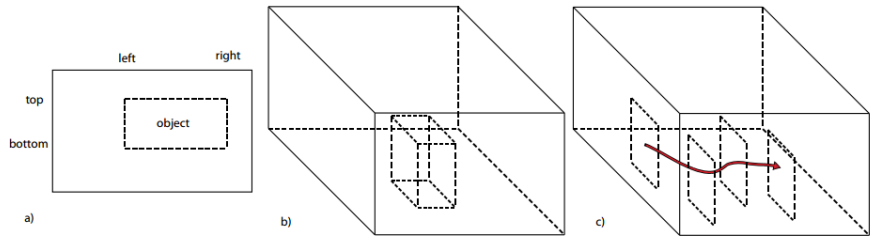


Figure 1: **Complexities of object and action localization**: a) Object localization is of O(n^4). b) Action localization by subvolume search is of O(n^6). c) Spatio-temporal action localization in a much larger search space.

15

[15] Du Tran and Junsong Yuan. "Max-margin structured output regression for spatio-temporal action localization". In: *NIPS*. 2012, pp. 350–358.

Tran et al. 2012 NIPS

Approach:

- **Finding an optimal spatio-temporal path** to detect and localize actions (A map from video X to the spatio-temporal path Y).
- Learn a **discriminant function** F : X*Y -> R, which is a **compatibility function** between X and Y.
- Train the function F to get the parameter w.
- When F is fixed, the prediction can be obtained by **maximization of the function**.
- The inference and training problem can be **solved by Max-Path algorithm**.

Tran et al. 2012 NIPS

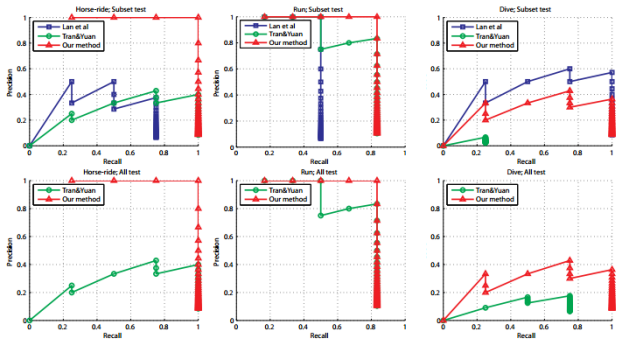


Figure 2: **Action detection results on UCF-Sport**: detection curves of our proposed method compared with [12] and [23]. Upper plots are detection results evaluated on subset frames given by [12], while lower plots are the results of all-frame evaluations. Except for diving, our proposed method significantly improves the other methods.

Eval. Set	Method	H-Ride	Run	Dive	Average
Subset	[12]	21.75	19.60	**42.67**	28.01
	[23]	62.19	50.20	16.41	42.93
	Our	**68.06**	**61.41**	36.54	**55.34**
All	[12]	N/A	N/A	N/A	N/A
	[23]	63.06	48.09	22.64	44.60
	Our	**64.01**	**61.86**	**37.03**	**54.30**

Table 1: **Action localization results on UCF-Sport**: comparisons among our proposed method, [12], and [23]. The upper section presents results evaluated on a subset of frames given by [12], while the lower section reports results from evaluating on all frames. Our method improves 27.33% from [12] and 12.41% from [23] on subset evaluations and improves 9.7% from [23] on all-frame evaluations. N/A indicates not applicable.

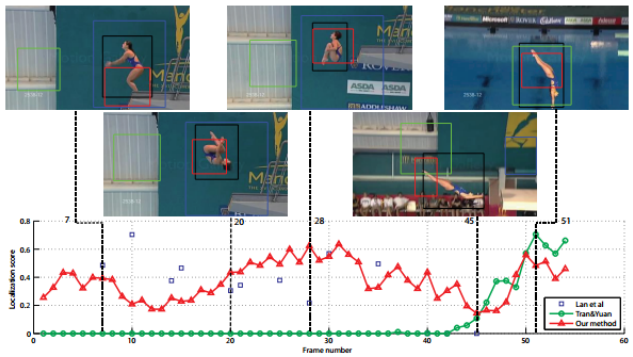


Figure 3: **Visualization of diving localization**: the plots of localization scores of different methods on a diving video sequence. Lan et al's [12] results are visualized in blue, Tran and Yuan's [23] are green, ours are red, and ground truth are black boxes. Best view in color.

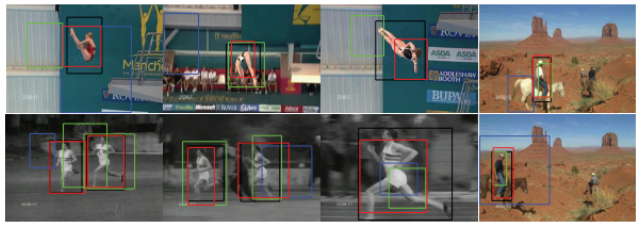


Figure 4: **Action detection and localization on UCF-Sport**: Lan et al's [12] results are visualized in blue, Tran and Yuan's [23] are green, ours are read, and ground truth are black. Our method and [23] can detect multiple instances of actions (two bottom left images).

Tran et al. 2012 NIPS

Comments

- A advanced idea that treat the localization as a regression problem, instead of classification problems, or a searching problem.
- An efficient way to solve the problem, by some theory.
- A new perspective to treat the action localization.
- Similar to metric learning network.

Lan et al. 2011 ICCV

Problems:

- **Joint action localization and recognition**, due to the opinion that action localization will benefit the action classification.

Motivation:

- Bags-of-words statistical representation **lack important cues** about **the spatial arrangement of features**, as well as **explicit modeling of the human figure**.
- General figure-centric representation relies on either a **template matching strategy** or **human detection and tracking** as input, which is unreliable in many situations.

16

[16] Tian Lan, Yang Wang, and Greg Mori. "Discriminative figure-centric models for joint action localization and recognition". In: *ICCV*. IEEE. 2011, pp. 2003–2010.

Lan et al. 2011 ICCV

Approach:

- Design a model which **combines the global statistical representation and figure-centric structural representation**.
- For avoiding the unrelible detection or tracking result, they **treat the position of human as latent variable** in a discriminative latent variable model.
- A scoring function to measure the compatibility between a video I, an action label y, and the configurations of bounding boxes L.
  - Unary potential: compatibility between y, li, zi in each frame Ii.
  - Pairwise potential: compatibility between two neighboring frames and assesses how likely they are to contain the same person.
  - Global action potential: compatibilty between action label y and a global feature vector of the whole video.

Lan et al. 2011 ICCV

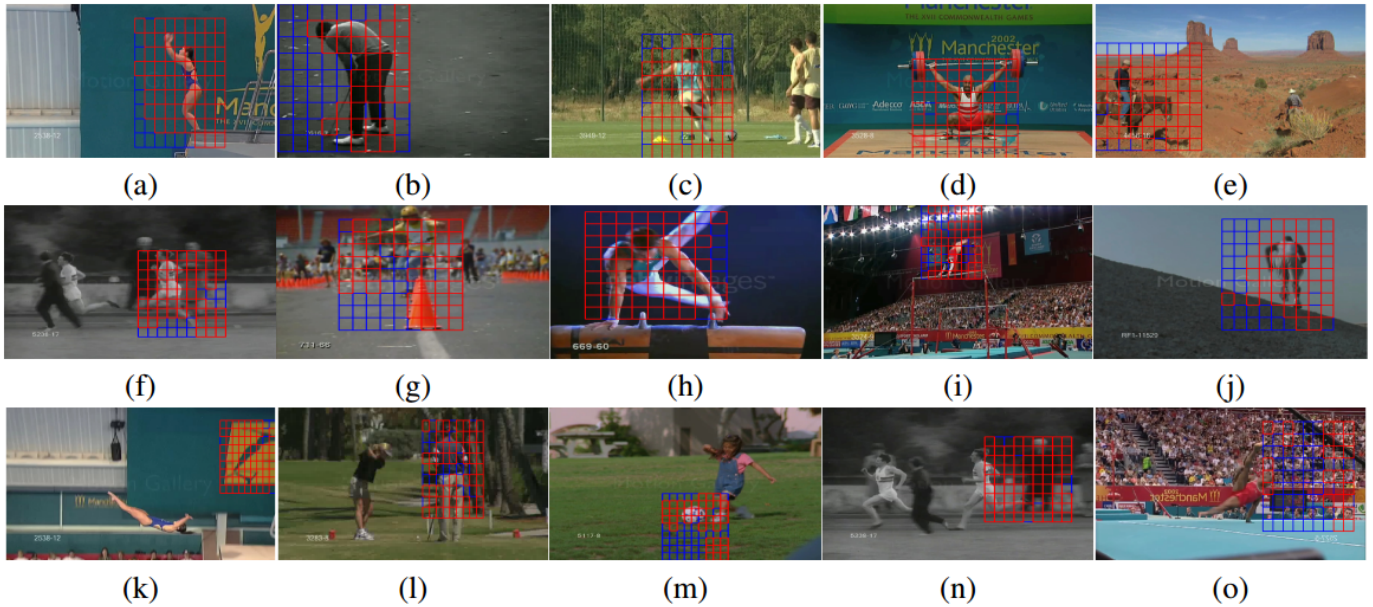


Figure 5: (Best viewed in color) Visualization of localization results and the learnt discriminative cells of each bounding boxes (or equivalently latent variables $\mathbf{z}$) for each action category. The red cells indicate the regions are inferred as discriminative regions by our model, the blue cells indicate the regions are not discriminative. The first two rows show correct examples. We can see that most of the discriminative regions (red cells) are on the person performing the action of interest or discriminative context such as the golf club, soccer ball, barbell, horse and bar in (b)-(e) and (i) respectively. The last row shows incorrect examples. We can see that most of the incorrect localizations are due to background clutter (e.g. the person shaped logo in (k)), or strong scene context (e.g. the soccer ball in (m)).

Lan et al. 2011 ICCV

Comments

- **A discriminative approach** for action localization, but **aware the detailed region**, which devoted in the action (Attention idea), making the approach more robust.
- Some **manual parameters**, hard to decide.
- Embed the **figure-centric z into the model**, the original features are still bags of general features.
- **Different areas in the frame definitely devote the results with different weights**.

Tran et al. 2011 CVPR

Problems:

- Search spatio-temporal video patterns with application like video event and human action detection.

Challenges:

- Most of the current spatio-temporal sliding window search methods **only support sliding windows of constrained structure, i.e., the 3-D bounding box**, which is not suitable for video localization.
- It becomes very **time consuming** to search 3D sliding windows.

1718

[17] Du Tran and Junsong Yuan. "Optimal spatio-temporal path discovery for video event detection". In: *CVPR*. IEEE. 2011, pp. 3321–3328.

[18] Du Tran, Junsong Yuan, and David Forsyth. "Video event detection: From subvolume localization to spatiotemporal path search". In: *IEEE transactions on pattern analysis and machine intelligence* 36.2 (2013), pp. 404–416.

Tran et al. 2011 CVPR

Main idea:

- Formulate the video event detection as a **spatio-temporal path discovery problem**, only modelling the center of the bounding box.

Approach:

- The problem is formulated as **finding the optimal path p* with the highest accumulated score**. (*Maximum path* problem)

$$p^* = \underset{p \in path(\mathcal{G})}{\operatorname{argmax}} M(p)$$

- Propose a new algorithm with **message passing mechanism** for the Max-Path discovery problem, with the complexity of only *O(whn)*.

Tran et al. 2011 CVPR

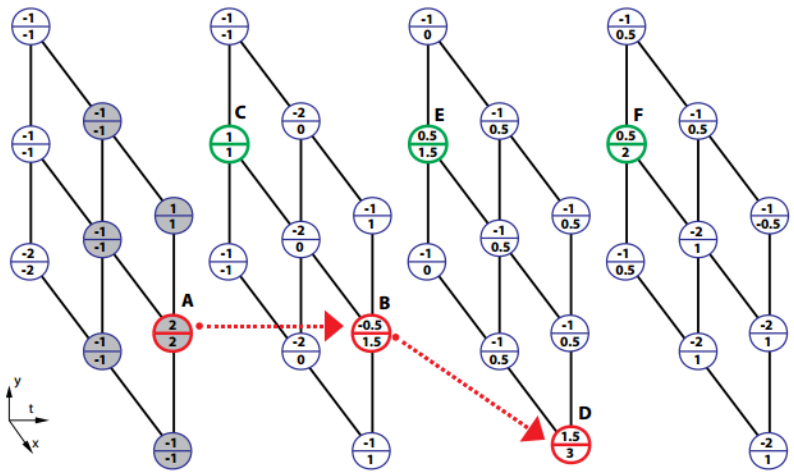


Figure 3. **A message passing example**: an example of Max-Path algorithm applied to a $3 \times 3 \times 4$ video. Each node is denoted with a local discriminative score (upper number), and the best accumulated score (lower number). In the first frame, all the best accumulated scores are initialized by their corresponding local discriminative scores. In the second frame, B can grow further from A which has the best accumulated score among B's neighbors (shaded nodes), while C needs to start a new path. The final best path is A-B-D (red nodes), and C-E-F is the second best path (green nodes).

Input: $M(u,t)$: the local discriminative scores;
Output: $S(u,t)$: the accumulated scores of the best path leads to (u,t);
$P(u,t)$: the best path record for tracing back;
S^* : the accumulated score of the best path;
l^* : the ending location of the best path;

```
begin
    S(u,1) = M(u,1), ∀u;
    P(u,t) = null, ∀(u,t);
    S* = −∞;
    l* = null;
    for i ← 2 to n do
        foreach u ∈ [1..w] × [1..h] do
            v0 ← argmax_{v∈N(u)} S(v, i − 1);
            if S(v0, i − 1) > 0 then
                S(u,i) ← S(v0, i − 1) + M(u,i);
                P(u,i) ← (v0, i − 1);
            else
                S(u,i) ← M(u,i);
            end
            if S(u,i) > S* then
                S* ← S(u,i);
                l* ← (u,i);
            end
        end
    end
end
```

Algorithm 1: Message forwarding algorithm

Tran et al. 2011 CVPR

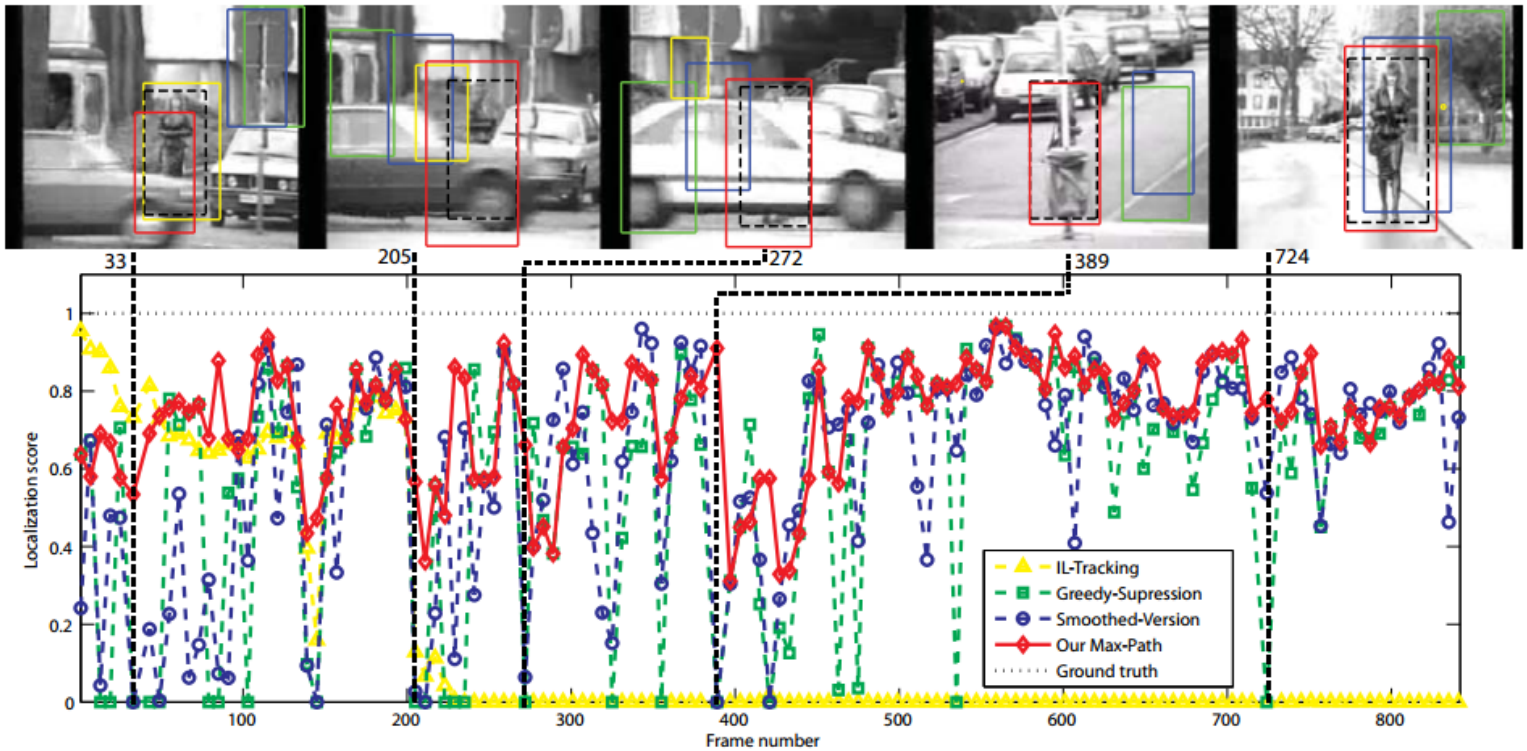


Figure 6. **Detection and localization results**: the plots of localization scores from different algorithms on an outdoor walking sequence with visualized snapshots. IL-Tracking [17] works only at the beginning, then loses the targets when occlusions occur. Greedy-suppression and smoothed-version perform poorly due to false positives and missed detections. Max-Path significantly outperforms the other algorithms with the globally optimized solution. The data points are dropped by ratio 1 : 7 for better representation (best viewed in color).

Tran et al. 2011 CVPR

Comments:

- **A smooth trajectory** can model the temporal continuity.
- **Simplify** the bounding box as a trajectory is a good way to **improve efficiency**.
- All the search problem has a assumption, that **the templates are quite similar with every frame in the action**. But in real scenes, it maybe failed.
- And the **spatial-temporal local features without structural information** is the best choice for features?

Gaidon et al. 2013 PAMI

Problems: (temporal action localization)

- Finding **if and when** an action is performed in a database of long and **unsegmented** videos.

Motivation:

- BOF models suffer two limitations: **orderless models**, and **segmented test videos**.
- A large number of actions can be naturally defined by **a composition of simpler temporal parts**.

Fig. 1: *Examples of actom annotations for two actions.*

Gaidon et al. 2013 PAMI

Approach:

- Model an action as **a small sequence of key atomic action units**, which we refer to as **actoms**. Usually 2-4 actoms represent an action. Make annotations.
- Propose a ASM classifier to train and perform a sliding central frame temporal localization with prior probability distribution.
- Classification by localization

Gaidon et al. 2013 PAMI

Experiments:

- Coffee and cigarettes
- DLSBF
- Hollywood2

Fig. 7: Frames of the top 5 actions localized with ASM for "Drinking" (top row) and "Open Door" (bottom row).

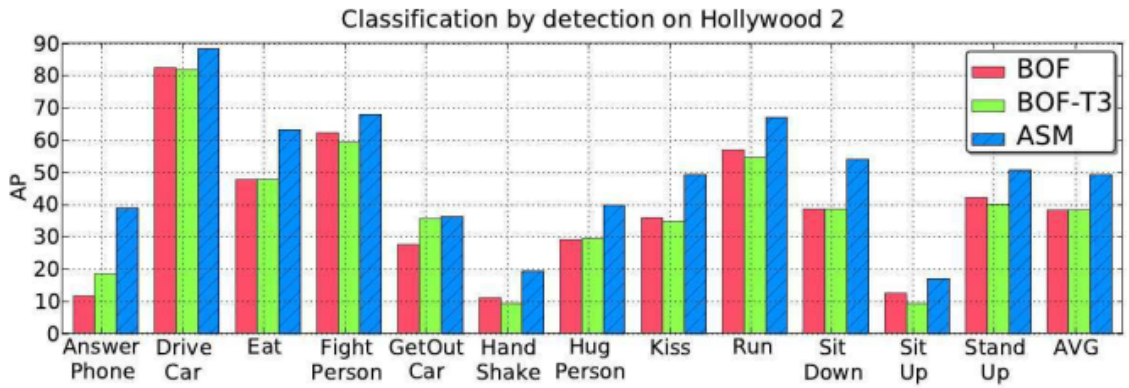


Fig. 8: Classification by localization results, in Average Precision (AP), on the "Hollywood 2" dataset [12]. "BOF" and "BOF-T3" are sliding-window approaches using BOF and its temporally structured extension. Our approach is "ASM". "AVG" contains the average performance over all classes (BOF: 38.1%, BOF T3: 38.2%, ASM: 49.3%).

Gaidon et al. 2013 PAMI

Comments:

- User-defined temporal parts description for action.
- Flexible and simplify the problem.
- Ignore the spatial information, the distinctive between foreground and background.
- Manual description is hard to describe more complex motion.

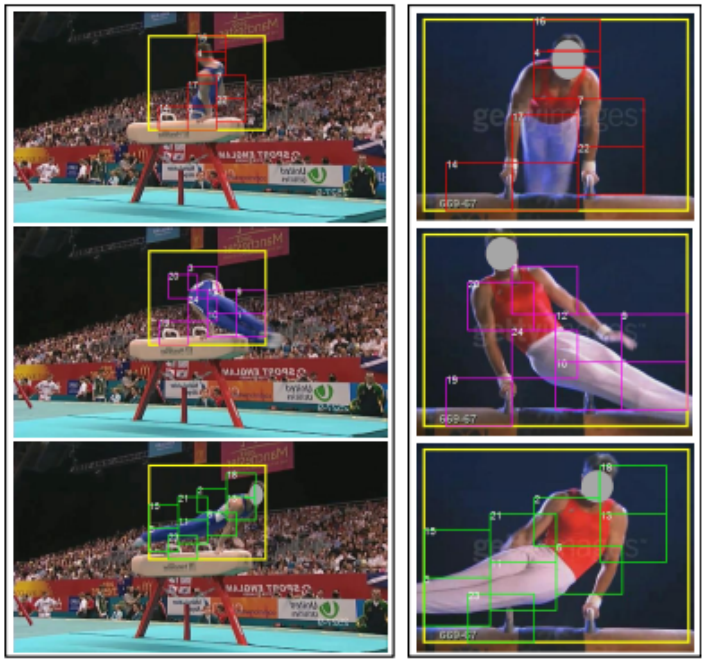
Tian 2013 CVPR

Problems:

- Perform DSP (deformable part model) to generalize for spatiotemporal representation.

Motivation

- Deformable parts can avoid the distraction of cluttered background.

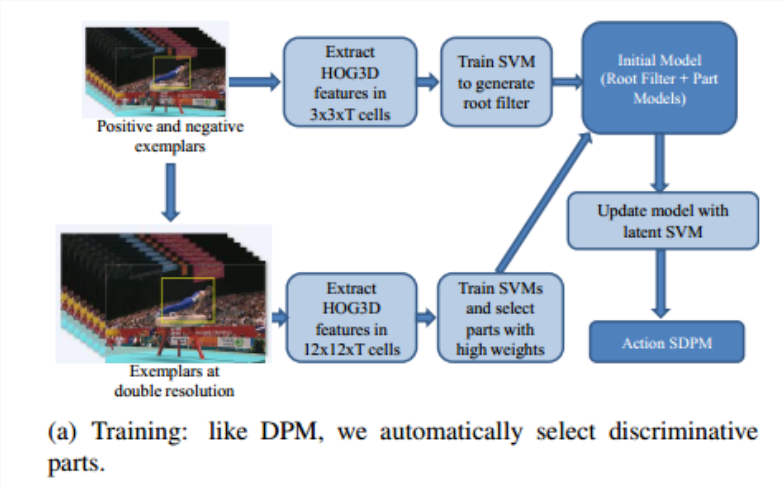


(a) training (b) testing

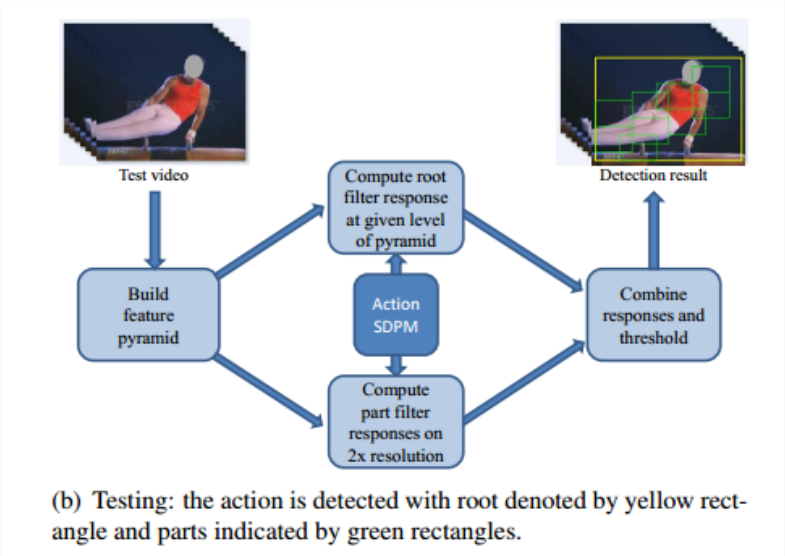
Figure 1. An example of "Swing Bench" SDPM (left) and its localization result in a test video from UCF Sports (right). This model consists of several parts across three temporal stages (middle frame of each stage shown in each row). The large yellow rectangle indicates the area under the root filter and the small red, magenta, and green ones denote parts. Although trained in videos with cluttered background at a different scale, the SDPM successfully localizes the target action in both space and time.

Tian 2013 CVPR

Approach:

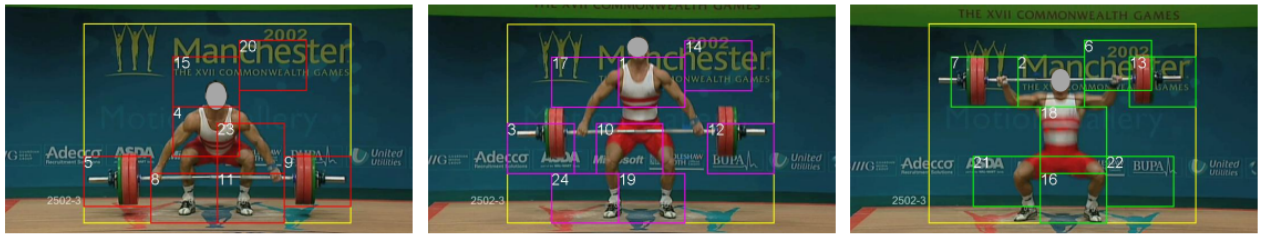


(a) Training: like DPM, we automatically select discriminative parts.



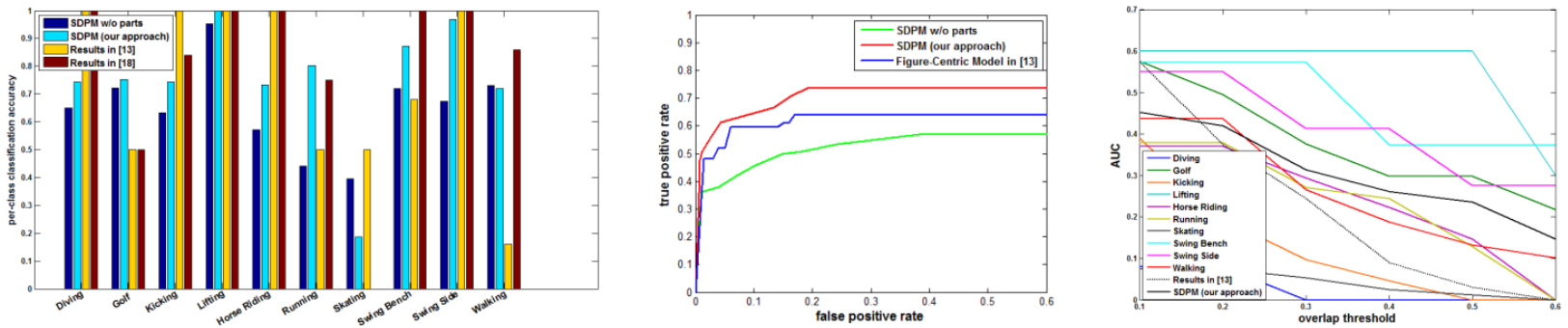
(b) Testing: the action is detected with root denoted by yellow rectangle and parts indicated by green rectangles.

Tian 2013 CVPR



(a) (b) (c)

Figure 4. SDPM for "lifting" in UCF Sports, with parts learned in each of the temporal stages. There are in total 24 parts for this SDPM and the index of each part is indicated at the left top corner of corresponding small rectangle. See Fig. 1 for example in clutter.



(a) (b) (c)

Figure 5. Direct comparisons on UCF Sports vs. [12, 17]. (a) classification; (b) detection, ROC at overlap threshold of $\theta = 0.2$; (c) detection, AUC for θ from 0.1 to 0.6. The black solid curve shows the average performance of SDPM and the black dotted curve shows the average performance of [12]. Other curves show SDPM results for each action. (Best viewed in color.)

Tian 2013 CVPR

Comments:

- Extend the 2D images to 3D spatiotemporal volumes.
- A reasonable idea to contribute different areas to the action detection results.

Jain et al. 2014 CVPR

Problems:

- Detect **when** and **where** an action of interest occurs.

Motivation:

- Investigate the **selective search sampling strategy** for videos.
- Incorporate motion information in various stages of the analysis with **independent motion evidence**.

21

[21] Mihir Jain et al. “Action localization with tubelets from motion”. In: *CVPR*. 2014, pp. 740–747.

Jain et al. 2014 CVPR

Action localization pipeline:

- Super-voxel segmentation;
- Iterative generation of additional tubelets;
- Descriptor computation;
- Classification step.

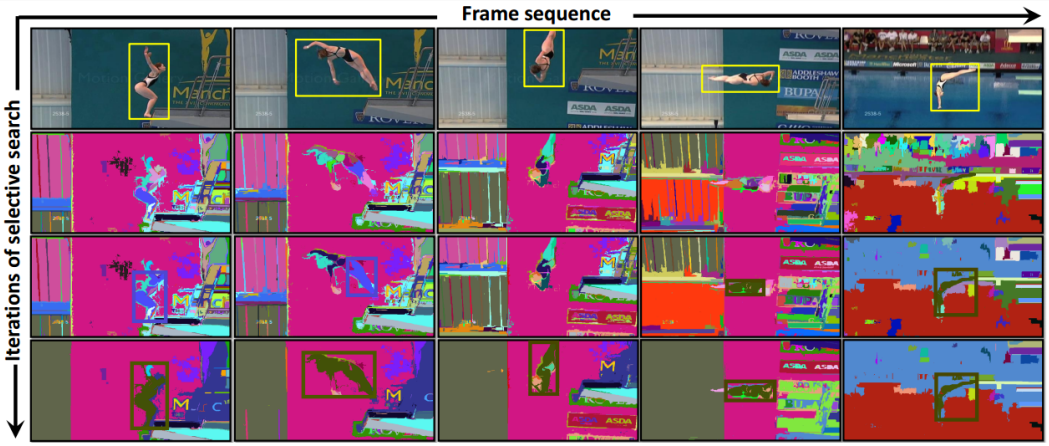
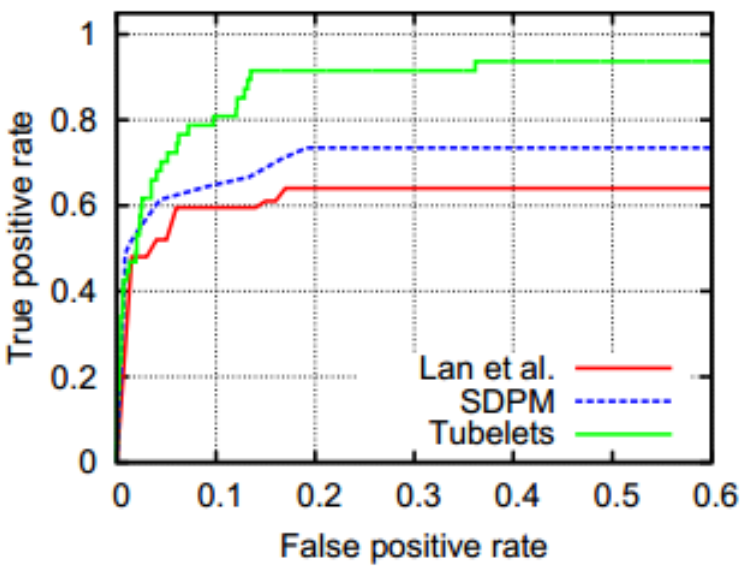


Figure 2. Illustration of hierarchical sampling of tubelets. *Top.* A sampled sequence of frames (1^{st}, 15^{th}, 25^{th}, 35^{th}, 50^{th}) associated with action 'diving' from UCF-Sports dataset. The yellow bounding boxes represent the ground-truth tubelet. *Row 2* shows the video segmentation used 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.

Jain et al. 2014 CVPR



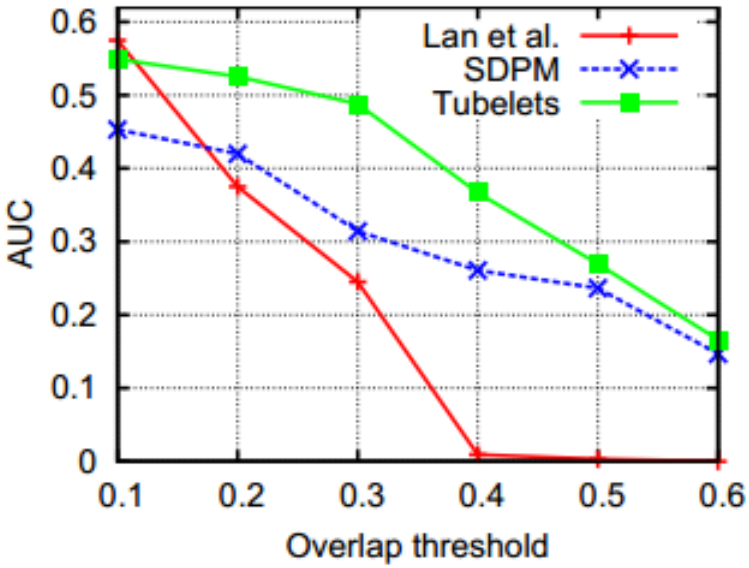


Figure 5. Comparison with concurrent methods [16, 24] on UCF-Sports: ROC at σ=0.2 and AUC for σ from 0.1 to 0.6.

Jain et al. 2014 CVPR

Comments:

- A try of transforming search selective from 2D image to 3D videos.
- A effective method that firstly pixel-wise locate the actor in the video, and then classify its class, is very efficient.
- Incorporate short-term motion features.

Ma et al. 2013 ICCV

Problems:

- Develop a representation of action for action recognition and localization.

Motivation:

- STIPs and trajectory only focus on non-static parts of the video.
- We argue that both non-static and relevant static parts are important.
  - Some static parts of the space-time video volume can be helpful in recognizing human actions.
  - Estimating the location of the action performer is also desired in addition to recognizing the action.

22

[22] Shugao Ma et al. "Action recognition and localization by hierarchical space-time segments". In: *ICCV*. 2013, pp. 2744–2751.

Ma et al. 2013 ICCV

Approach:

- Propose a representation that called **hierarchical space-time segments**, which is organized in a two-level hierarchy.
  - Root space-time segments that may contain the whole human body;
  - Space-time segments that contain parts of the root.

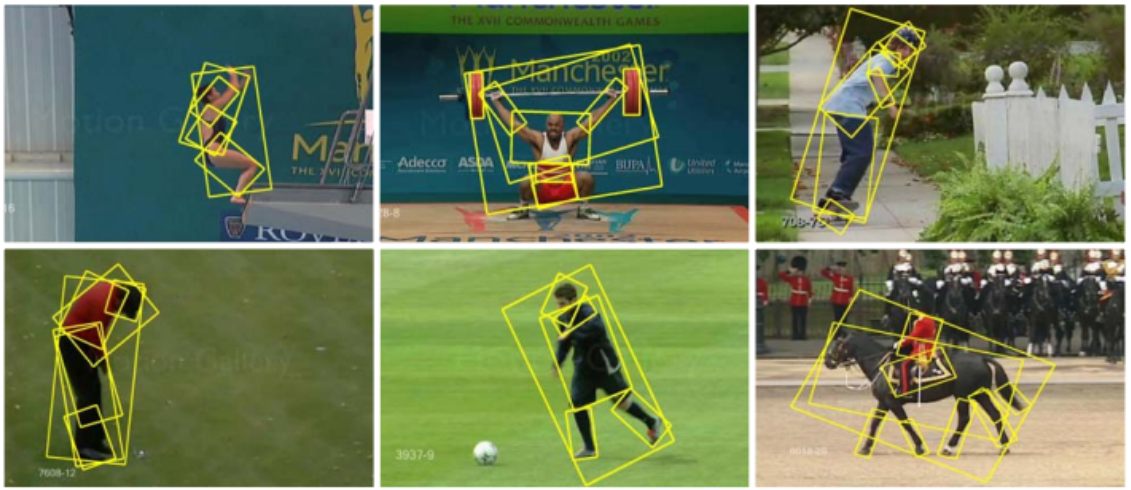


Figure 1. Extracted segments from example video frames of the UCF Sports dataset. Yellow boxes outline the segments. Boxes within a box indicate child-parent relationships.

Ma et al. 2013 ICCV

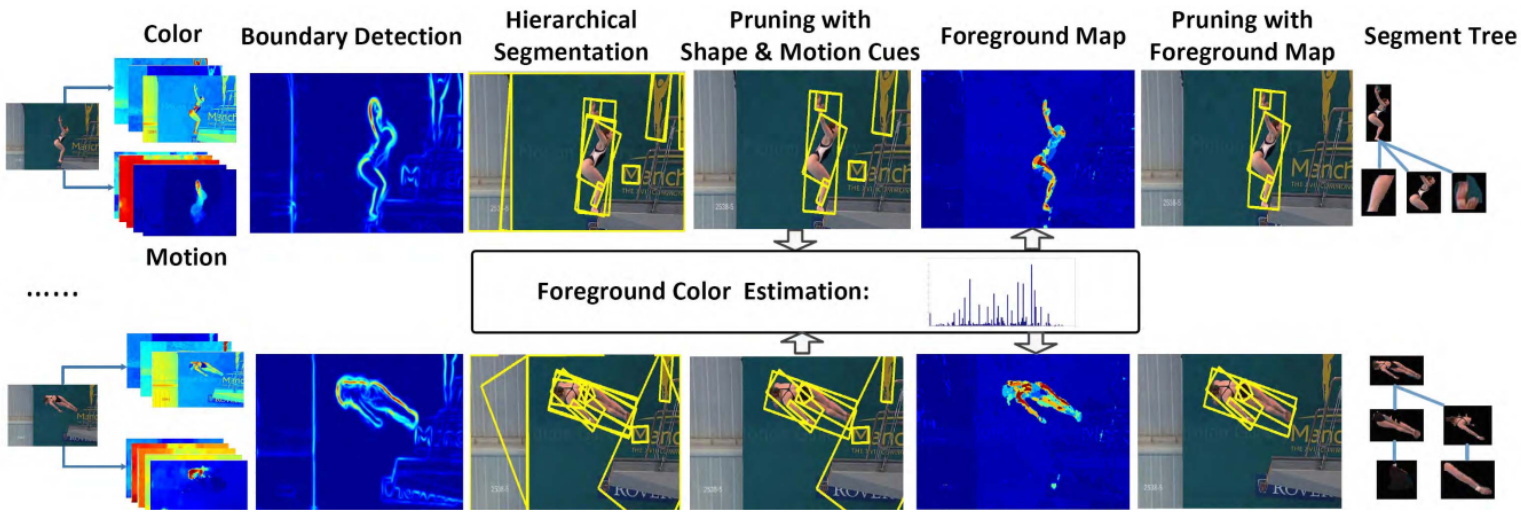


Figure 3. The pipeline for hierarchical video frame segments extraction.

Ma et al. 2013 ICCV

- Hierarchy structure to model the human body, providing more information for the recognition, as well as an approach of localizing.
- Highly dependent on the quality of multiple segment with tracking.
- Ignore the information of background?
- Bags-of-features? Any other replacement?

Oneata et al. 2014 CVPR

Motivation:

- **Fisher vector(FV) image representation**, a high-dimensional extension of the popular bag-of-word representation is very effective in image task.
- However, it leads to **large storage requirements**.
- **Normalized FV prevent** the use of integral image techniques to efficiently **aggregate local features or scores**.

Contribution:

- **Approximate FV normalization**, relying on pre-computed cumulative sums of local visual word assignments, scores, and l2 norm.
- The approximations the normalized FV becomes amenable to efficient localization using **branch-and-bound search**.

23

[23] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. "Efficient action localization with approximately normalized fisher vectors". In: *CVPR*. 2014, pp. 2545–2552.

Oneata et al. 2014 CVPR

Power norm.	ℓ_2 normalization	Hollywood2	HMDB
No	No	55.2	43.1
Exact	No	62.0	51.7
No	Exact	60.1	46.8
Exact	Exact	62.4	52.2
Approximate	Exact	62.1	52.1
Approximate	Approximate, $n = 5$	60.1	52.6
Approximate	Approximate, $n = 10$	60.2	52.4
Approximate	Approximate, $n = 20$	60.2	52.6
Approximate	Approximate, $n = 40$	60.6	52.5
Approximate	Approximate, $n = 80$	60.7	52.2
Approximate	Approximate, $n = 160$	61.1	52.2
Wang *et al*. 2013 [31]		59.9	48.3
Oneata *et al*. 2013 [21]		61.9	51.9
Jain *et al*. 2013 [13]		62.5	52.1
Wang *et al*. 2013 [32]		64.3	57.2

Table 1. Action classification performance. For the ℓ_2 approximation we evaluate using cells of n frames, for $n = 5$ to $n = 160$.

Oneata et al. 2014 CVPR

- Successfully transform the Fisher Vector into the video problems, and propose an efficient algorithm to approximate the result.
- Branch-and-bound searching problem.
- Feature extraction?
- Temporal modelling?

Wang et al. 2014 ECCV

Motivation:

- Address the importance of **key pose** and **key motion** in understanding human actions.
  - Pose: Static configurations and geometric constraints of human body parts;
  - Motion: Local articulated movements of body parts and global rigid kinematics.

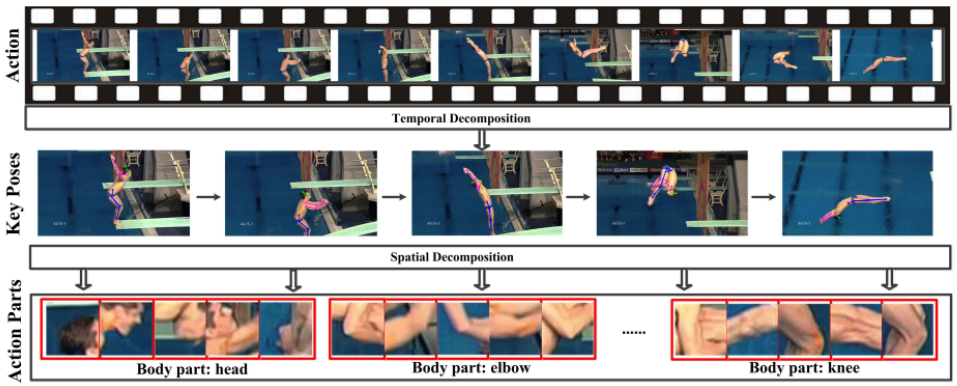


Fig. 1. Illustration of action decomposition. A video sequence first can be temporally decomposed into several short snippets, each of which corresponds to a key pose. For each key pose, the action can then be further decomposed spatially into several action parts (red boxes), each of which describes the appearance and motion of body part in a specific configuration. A body part is described by multiple action parts. Best view in color.

Wang et al. 2014 ECCV

Challenges:

- How to discover a collection of tightly-clustered action parts from videos?
- How to model the spatiotemporal relations of action parts?

Approaches:

- Annotate articulated human poses in the training data.
- Cluster cuboids that share similar pose configuration and motion patterns into consistent action parts, which we called **dynamic-poselets**.
- Propose a relational model, called sequential skeleton model(SSM), that is jointly learn the composites of mixture dynamic-poselets.
- Formulate the model learning problem in a structured SVM framework and use the dual coordinate-descent solver for parameter optimization.

Wang et al. 2014 ECCV

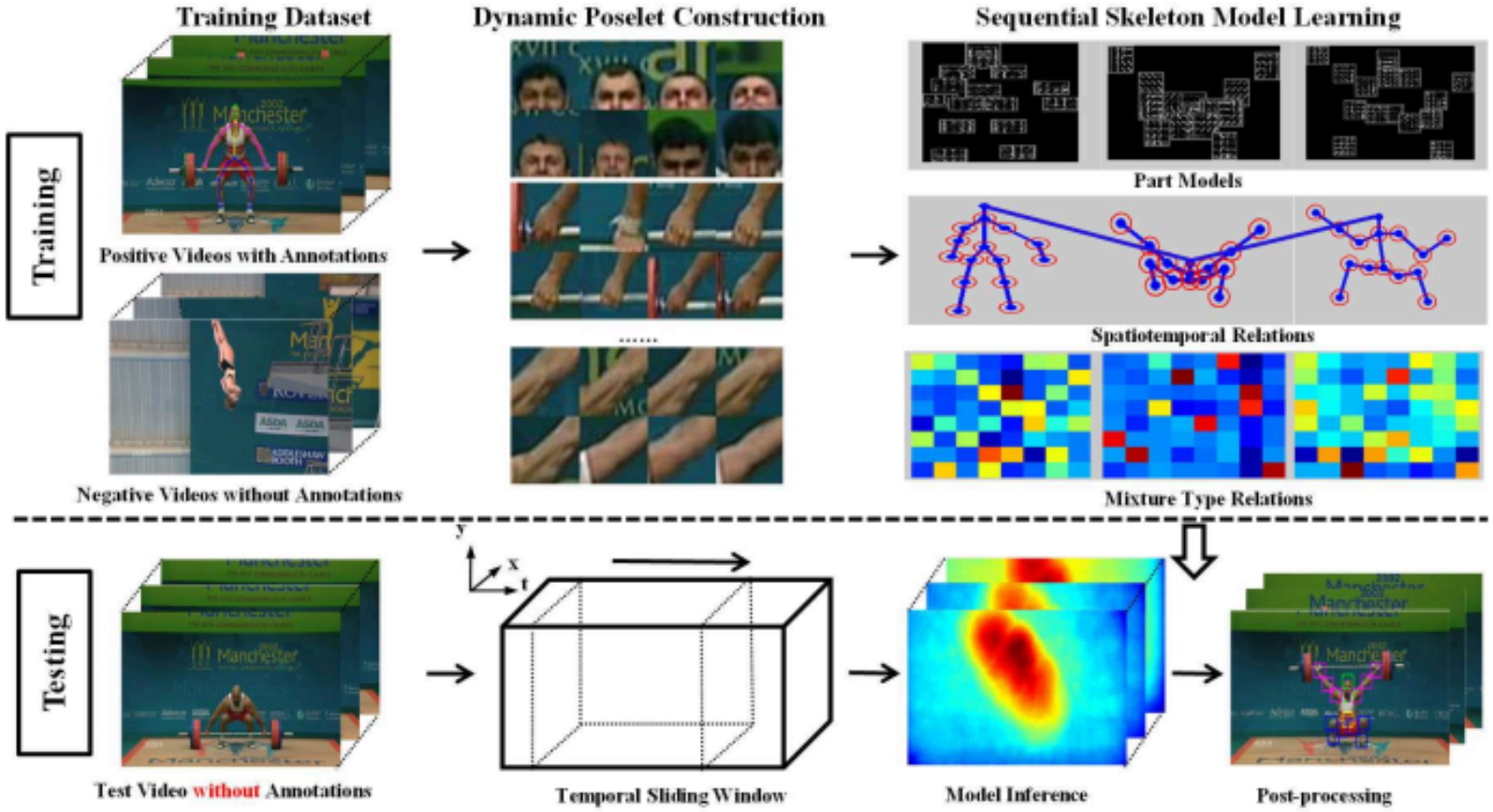


Fig. 3. Overview of our approach. For training, we annotate human joints for several key poses in the positive samples. We first cluster the cuboids around each human joint into dynamic-poselets. Then, each dynamic-poselet acts as a mixture of body parts and is fed into the SSM training. Our SSM is composed of three components: part models, spatiotemporal relations, and mixture type relations. For testing, we first use a temporal sliding window and then conduct inference of SSM. Finally, we resort to post-processing techniques such as no-maximum suppression to obtain the detection

Wang et al. 2014 ECCV

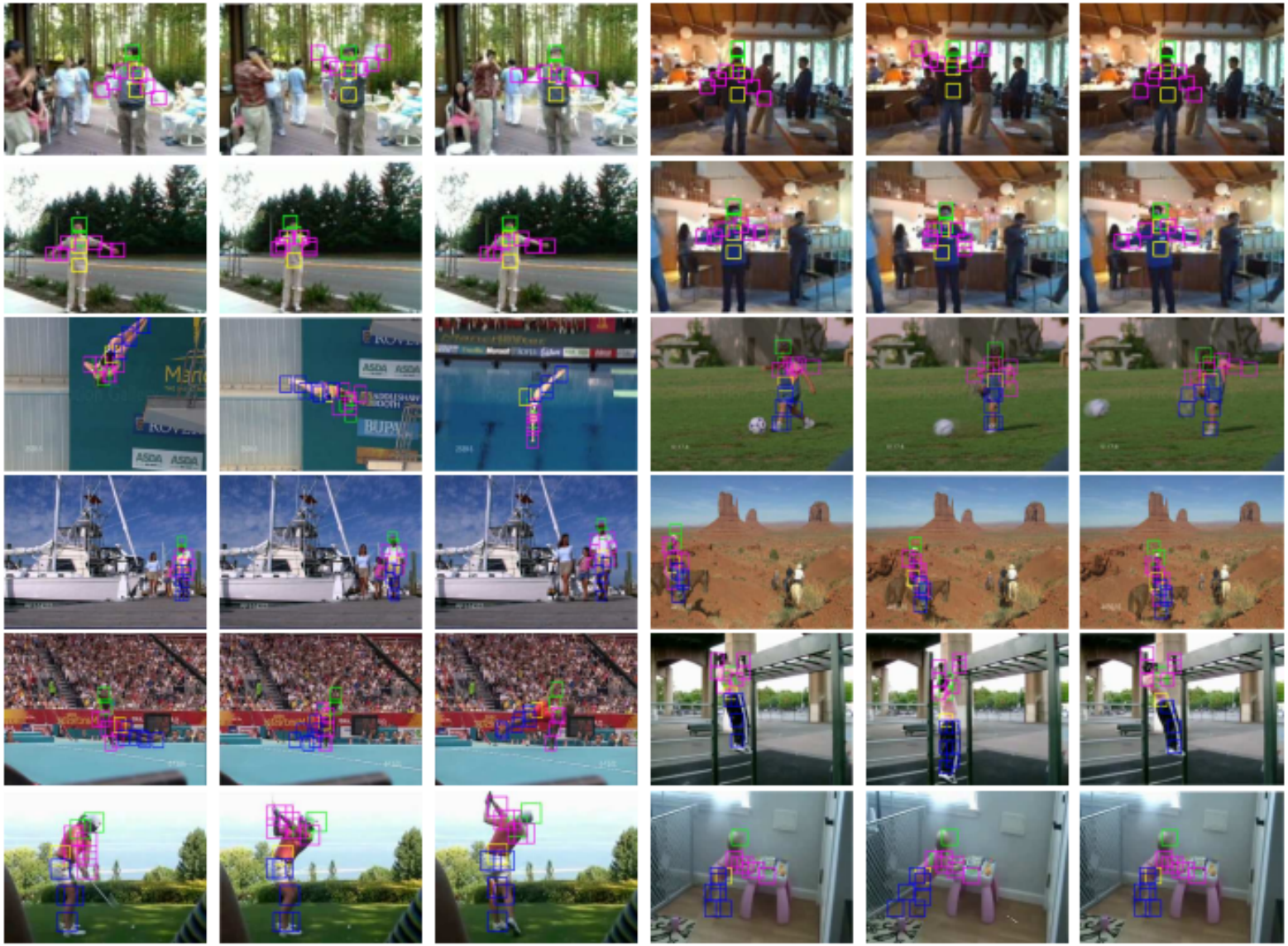


Fig. 7. Examples of action detection in three datasets. Our model is able to detect human actions and also estimate human poses accurately in most cases. Best viewed

Wang et al. 2014 ECCV

- An approach specially models the body parts in action videos.
- Guided by finer details on the human and sparse key temporal information, the results can be improved.
- Explicitly annotation. Small segments, reliable? reasonable? Engineering?

Yu et al. 2015 CVPR

Problems:

- Generate generic action proposals in unconstrained videos.

Challenges:

- Difficult to learn the actionness measure that can well differentiate human actions from the background clutters and other dynamic motions.
- The candidate number of action proposals can be much larger than that of the object proposals.

25

[25] Gang Yu and Junsong Yuan. "Fast action proposals for human action detection and search". In: *CVPR*. 2015, pp. 1302–1311.

Yu et al. 2015 CVPR

Main approach:

- Perform human and motion detection to generate candidate bounding boxes. (actionness score).
- Utilize the max sub-path search algorithm to locate the top-N maximal spatio-temporal paths based on ”actionness” score.
- To select right paths among overlap paths, we further formulate it as a maximum set coverage problem, with each bounding box as an element.

Algorithm 1 Action Proposal

Input: bounding box score $w(b_t^i)$,
Output: action candidates $\mathbf{P} = \{\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \cdots, \mathbf{p}^{(K)}\}$

$f_k = 0, b^{(k)} = \emptyset, \quad k = 1, 2, \cdots N$
for $t = 1 \rightarrow T$ **do**
  for $i = 1 \rightarrow N_b^t$ **do**
    $f(b_t^i) = \max_{b_{t-1}^j} \{f(b_{t-1}^j) + w(b_t^i), 0\}$ as Eq. 7
    if $f(b_t^i) > f_N$ **then**
      $f_N = f(b_t^i), \quad b^{(N)} = b_t^i$
    end if
  end for
end for
back trace to obtain $\mathbf{p}_i, i = 1, \cdots, N$
$k = 1, \quad \mathbf{P} = \emptyset$
repeat

Yu et al. 2015 CVPR

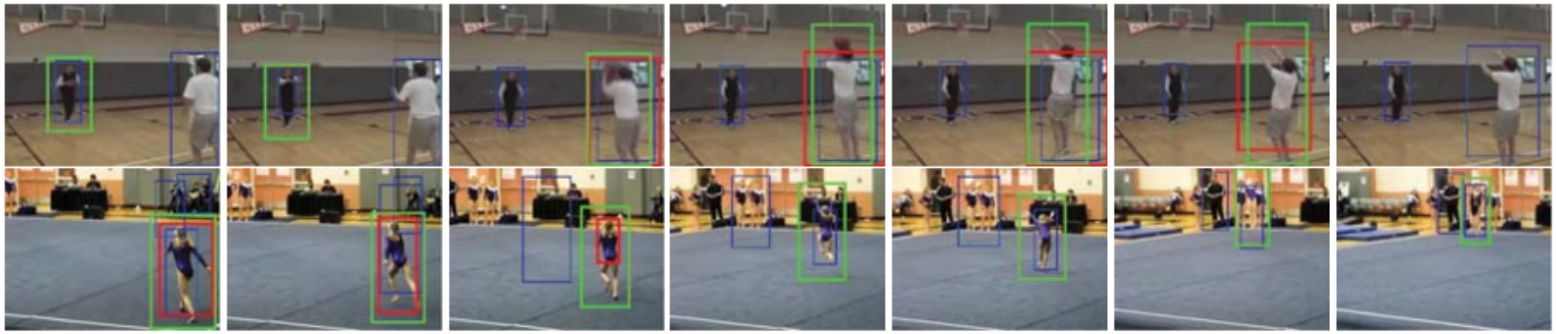


Figure 6. Action proposals (marked with blue rectangle) on two videos from UCF 101 dataset. The action proposal with maximum actionness path score $f(b^*)$ is marked with red rectangle and the ground-truth action is marked by green bounding box.

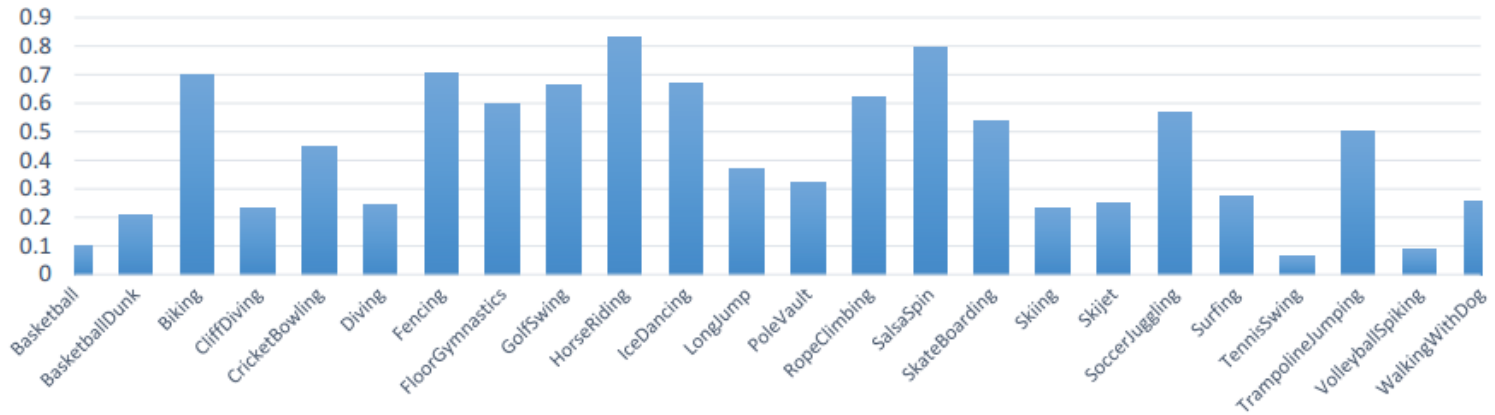


Figure 7. Our action detection results for UCF101 dataset based on average precision.

Yu et al. 2015 CVPR

Comments:

- Only human and motion score are hard to find the start and the end of the action for general proposal.
- And the definition of action is ambiguous. How to do action proposal? It's really **wired**.

Gkioxari et al. 2015 CVPR

Motivation:

- Inspired by the recent advances in the field of object detection, we start by selecting candidate regions and use CNN to classify them.

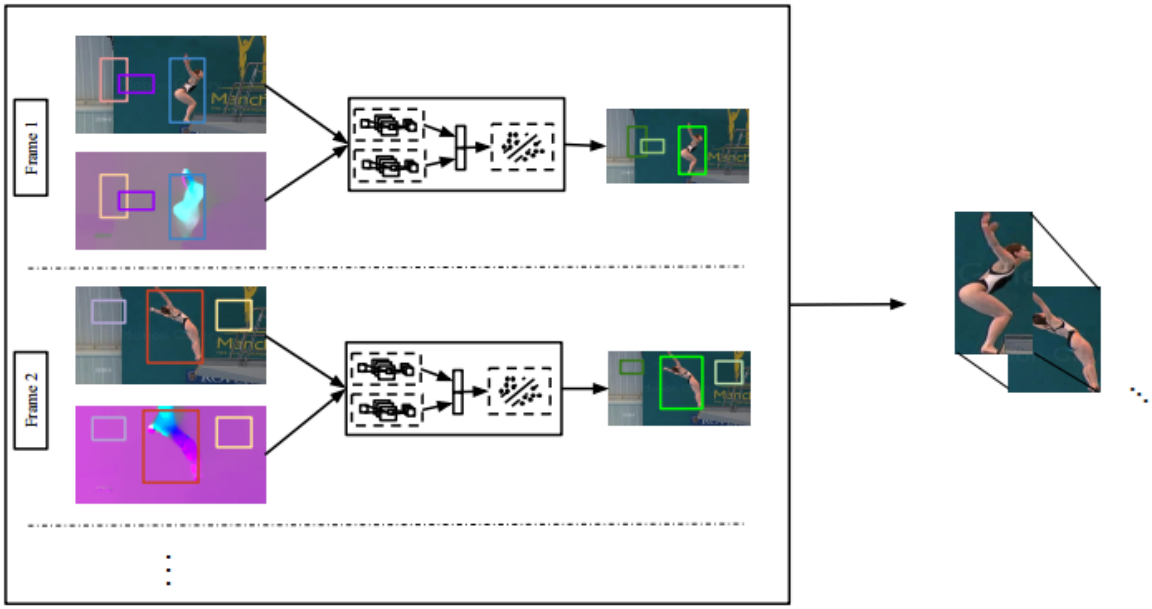
Approach:

- 1st: produce prediction of each object in every frame in spatio-temproal space, based on spatio-temporal features extracted by two-stream CNN.
- 2st: link detection to obtain action tubes, a.k.a optimal path recovery.

26

[26] Georgia Gkioxari and Jitendra Malik. "Finding action tubes". In: *CVPR*. 2015, pp. 759–768.

Gkioxari et al. 2015 CVPR



(a) Action detection at every frame of the video

(b) Linked detections in time produce *action tubes*

Figure 1: An outline of our approach. (a) Candidate regions are fed into action specific classifiers, which make predictions using static and motion cues. (b) The regions are linked across frames based on the action predictions and their spatial overlap. *Action tubes* are produced for each action and each video.

Gkioxari et al. 2015 CVPR

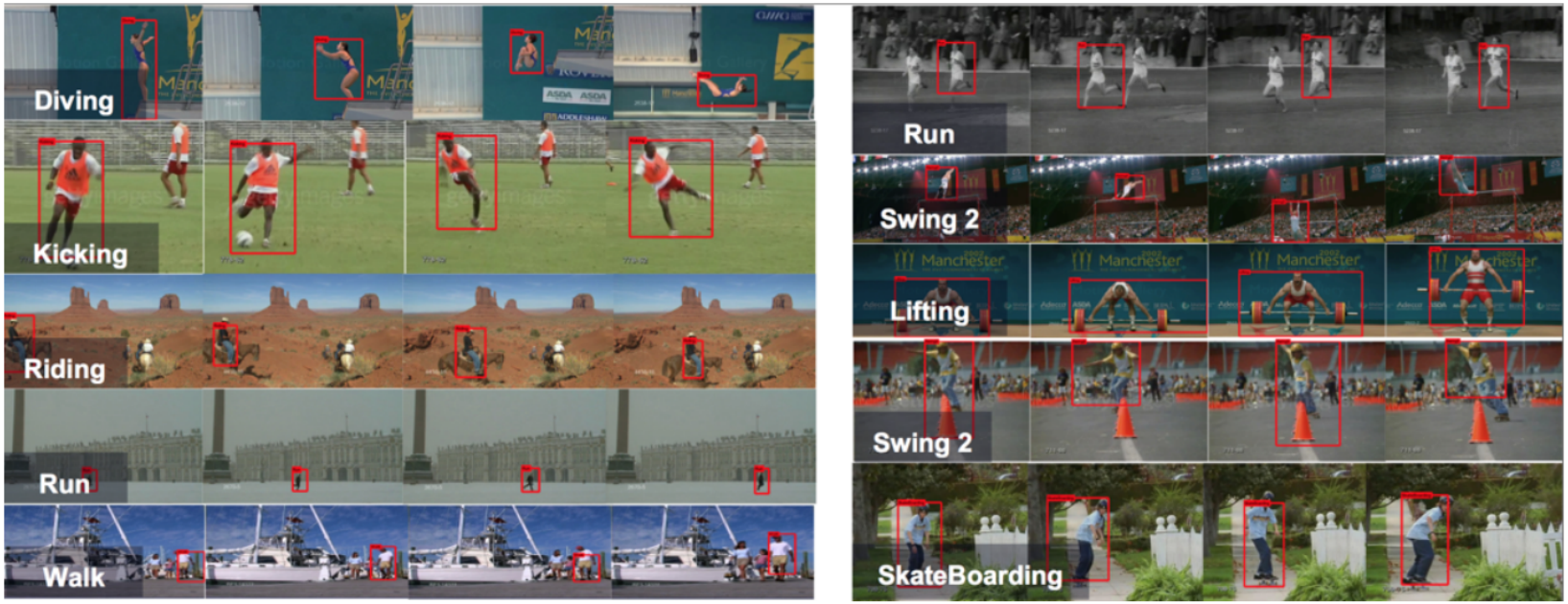


Figure 7: Examples from UCF Sports. Each block corresponds to a different video. We show the highest scoring action tube detected in the video. The red box indicates the region and the predicted label is overlaid. We show 4 frames from each video. The top example on the right shows the problem of tracking, while the 4th example on the right is a wrong prediction, with the true label being *Skate Boarding*.

Gkioxari et al. 2015 CVPR

Comments:

- Decompose the problem into individual frame prediction and link the action tube.
- Ignore the camera motion? Address the tracking problem?
- Utilize the information the static and kinematic cues.
- ”First predict, then link” VS ”first link, then predict”? Which one is better?

Weinzaepfela et al. 2015 ICCV

Spatio-temporal localization problems:

- Accommodate the uncertainty of per-frame spatial localization and the temporal consistency.

Motivation:

- **Extracting and scoring frame-level proposals**, with CNN descriptors based on appearance and motion information.
- **Track best candidates** based on a tracking-by-detection approach combining an instance-level and class-level detector.
- Score the tracks with the CNN features as well as **a spatio-temporal motion histogram descriptor** (STMH), which captures **the dynamics of an action**.
- **Temporal localization** is performed using a multi-scale sliding-window approach at the track level.

27

[27] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. "Learning to track for spatio-temporal action localization". In: *ICCV*. 2015, pp. 3164–3172.

Weinzaepfela et al. 2015 ICCV

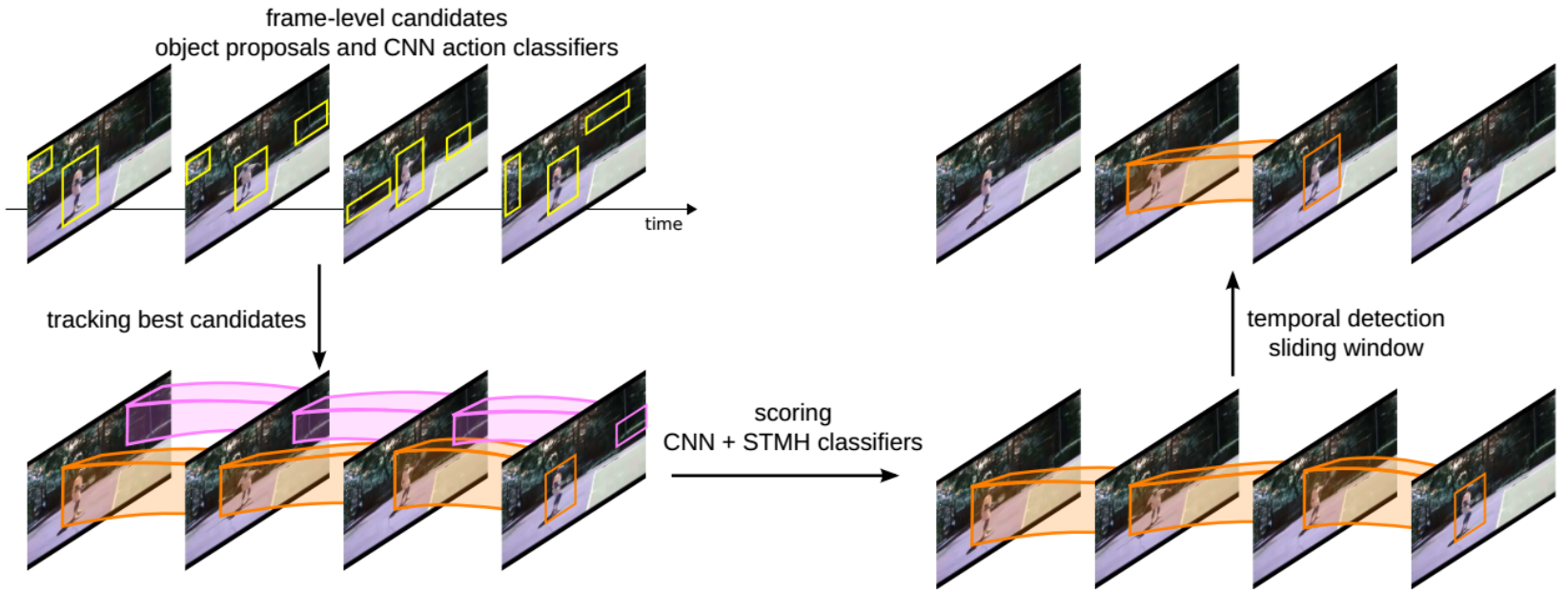


Figure 1. Overview of our action localization approach. We detect frame-level object proposals and score them with CNN action classifiers. The best candidates, in term of scores, are tracked throughout the video. We then score the tracks with CNN and spatio-temporal motion histogram (STMH) classifiers. Finally, we perform a temporal sliding window for detecting the temporal extent of the action.

Weinzaepfela et al. 2015 ICCV

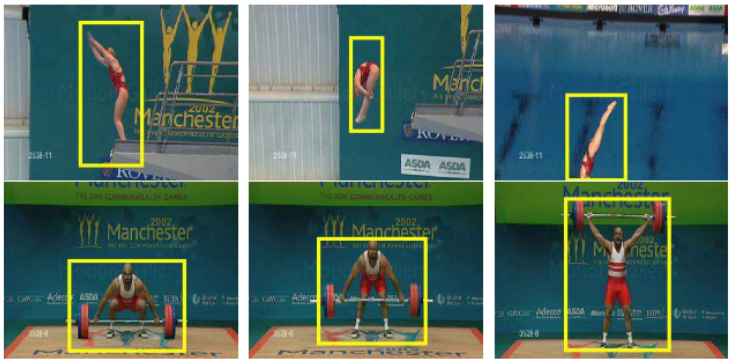


Figure 7. Example results from the UCF-Sports dataset.

δ	0.2	0.3	0.4	**0.5**
[11]				53.3
no STMH	58.1 $\pm$ 2.1	58.0 $\pm$ 1.9	57.7 $\pm$ 2.1	56.5 $\pm$ 2.6
ours	63.1 $\pm$ 1.8	63.5 $\pm$ 1.8	62.2 $\pm$ 1.9	60.7 $\pm$ 2.7

Table 4. Comparison to the state of the art on J-HMDB using mAP for varying IoU thresholds δ. We also report the standard deviation among the splits.

δ	0.05	0.1	**0.2**	0.3
[46]	42.8			
ours	54.28	51.68	46.77	37.82

Table 5. Localization results (mAP) on UCF-101 (split 1) for different IoU thresholds δ.

Weinzaepfela et al. 2015 ICCV

Comments:

- The first three steps aim to locate the action, where the first step locate the human, 2th step link the proposal, and 3th step pick the highest action proposal.
- Temporal localization is regarded as a classification problem.
- The localization of human is quite accurate, where we can take advantage of it for spatio-temporal action detection.

Chen et al. 2015 ICCV

Motivation:

- Understand the role that **human detection** (implicit or ex) can play in action detection.
- Leverage the fact that **action requires intentional motion**.

Main idea:

- Quantitatively analyze the proprieties that dense trajectories exhibit in space-time video regions of explicit intentional motion, i.e., regions where a human is performing an action.
- Found they're **consistent**.

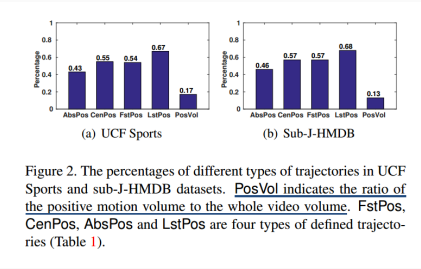


Figure 2. The percentages of different types of trajectories in UCF Sports and sub-J-HMDB datasets. PosVol indicates the ratio of the positive motion volume to the whole video volume. FstPos, CenPos, AbsPos and LstPos are four types of defined trajectories (Table 1).

[28]

[28] Wei Chen and Jason J Corso. "Action detection by implicit intentional motion clustering". In: *ICCV*. 2015, pp. 3298–3306.

Chen et al. 2015 ICCV

Approach:

- Generate a space-time trajectory graph.
- Implicit intentional motion clustering.
- Action Detection-by-Recognition.

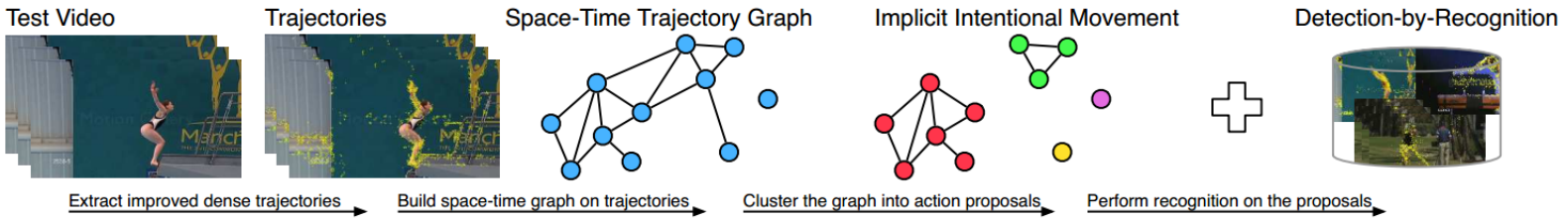


Figure 1. Illustration of our method. Given trajectories in a testing video, the spatio-temporal trajectory graph is used to select action proposals based on our notion of implicit intentional movement. Each cluster on the graph gives rise to an action proposal. The action classifier trained by videos for the action recognition task can be used to achieve action detection on these proposals, in our action detection-by-recognition framework.

Chen et al. 2015 ICCV

Chen et al. 2015 ICCV

Comments:

- Even though the analysis show that there are some consistence between trajectories and the action motion regions, but the no evidence that the the trajectories are all beneficial for the action classification.
- It's a weakly supervised method (no bounding box label) to locate the actor and motion regions, inspired by the progress of trajectories cluster in segmentation.
- The trajectory information is definitely beneficial for the location of action, but the trajectory features may be not discriminative for the action classification. Or through the extraction of features based on trajectory regions?
- Personally, I think the work is still focus on the localization process.

Marian et al. 2015 ICCV

Problems:

- Automatic extraction of foreground objects from videos.

Cues:

- Appearance-based static ”objectness” (selective Search);
- Motion information (dense trajectories);
- Transductive learning (detectors are forced to ”overfit” on the unsupervised data used for training)

Figure 1. Optical Flow between two consecutive frames can be used as a ”voting” mechanism for matching Bonding Boxes. The blue lines are dense trajectories in common between the two boxes, while the red lines are trajectory starting from the first box but not included in the second.

29

[29] Mihai Marian Puscas et al. “Unsupervised tube extraction using transductive learning and dense trajectories”. In: *ICCV*. 2015, pp. 1653–1661.

Yao et al. 2015 ICCV

Motivation:

- Use recurrent neural networks to model the dynamic temporal structure and produce natural language description for video description.

Figure 1. High-level visualization of our approach to video description generation. We incorporate models of both the local temporal dynamic (i.e. within blocks of a few frames) of videos, as well as their global temporal structure. The local structure is modeled using the temporal feature maps of a 3-D CNN, while a temporal attention mechanism is used to combine information across the entire video. For each generated word, the model can focus on different temporal regions in the video. For simplicity, we highlight only the region having the maximum attention above.

30

[30] Li Yao et al. "Describing videos by exploiting temporal structure". In: *ICCV*. 2015, pp. 4507–4515.

Peng et al. 2016 ECCV

Motivation:

- Extend Faster-RCNN to the action detection, i.e. frame-level based action detection.

Approach:

- Combined motion feature (optical flow input) for RPN to generate proposals.
- Multi-region scheme to improve the performance.

31

[31] Xiaojiang Peng and Cordelia Schmid. "Multi-region two-stream R-CNN for action detection". In: *ECCV*. Springer. 2016, pp. 744–759.

Peng et al. 2016 ECCV

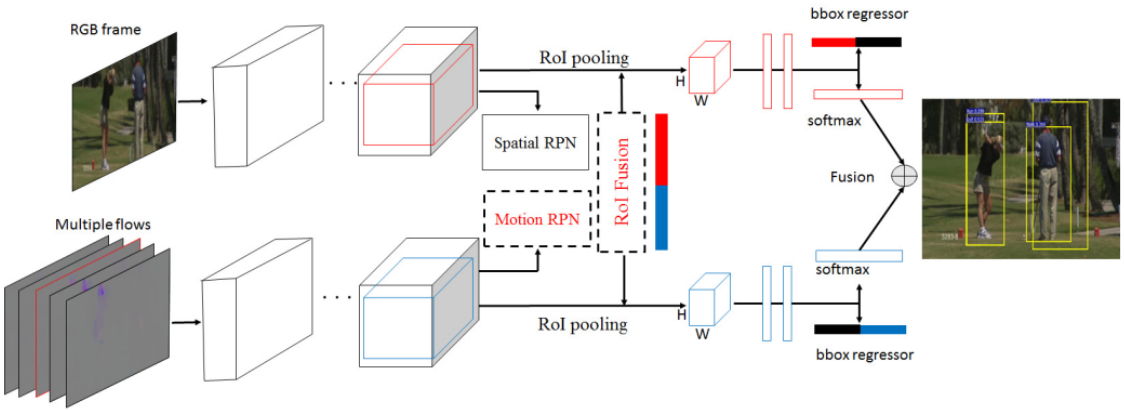


Fig. 1: Overview of our two-stream faster R-CNN.

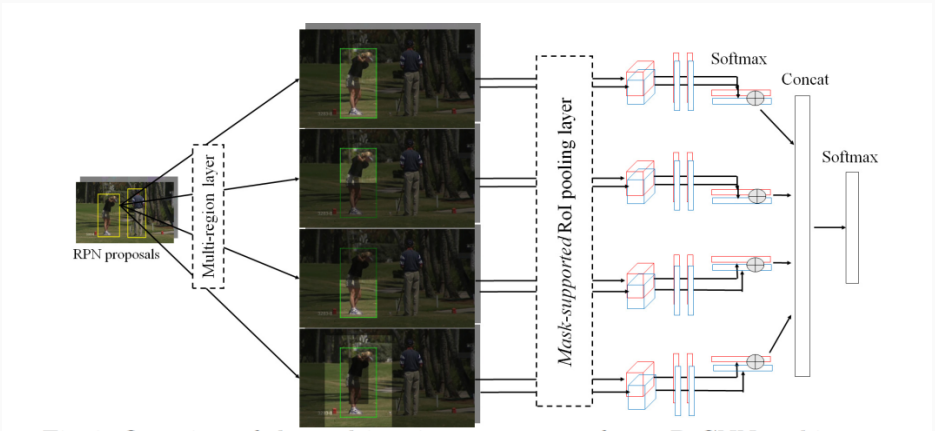


Fig. 3: Overview of the multi-region two-stream faster R-CNN architecture.

Peng et al. 2016 ECCV

Table 6: Comparison to the state of the art on three datasets. The IoU threshold δ for frame-mAP is fixed to 0.5.

		UCF-Sports		J-HMDB		UCF101 (split 1)			
	δ	0.2	0.5	0.2	0.5	0.05	0.1	0.2	0.3
video-mAP	Gkioxari *et al.* [8]	-	75.8	-	53.3	-	-	-	-
	Weinzaepfel *et al.* [9]	-	90.5	63.1	60.7	54.3	51.7	46.8	37.8
	Yu *et al.* [44]	-	-	-	-	49.9	42.8	26.5	14.6
	Our TS R-CNN	94.8	**94.8**	71.1	70.6	54.1	49.5	41.2	31.1
	Our MR-TS R-CNN	**94.8**	94.7	**74.3**	**73.1**	**54.5**	50.4	42.3	32.7
frame-mAP	Gkioxari *et al.* [8]	68.1		36.2		-			
	Weinzaepfel *et al.* [9]	71.9		45.8		35.84			
	Our TS R-CNN	82.3		56.9		**39.94**			
	Our MR-TS R-CNN	**84.5**		**58.5**		39.63			

Peng et al. 2016 ECCV

Comments:

- A straight and simple extension of fast R-CNN, but treat each frame individual, ignore the relationship between each frame.
- Thus cannot handle the occlusion, or other incontinent condition.
- And one question is, every frame in each action can be distinguished? Maybe ambiguous? Maybe need the more temporal information?

Wang et al. 2016 CVPR

Problems:

- Actionness estimation: quantify the likelihood of containing a generic action instance at a specific location.

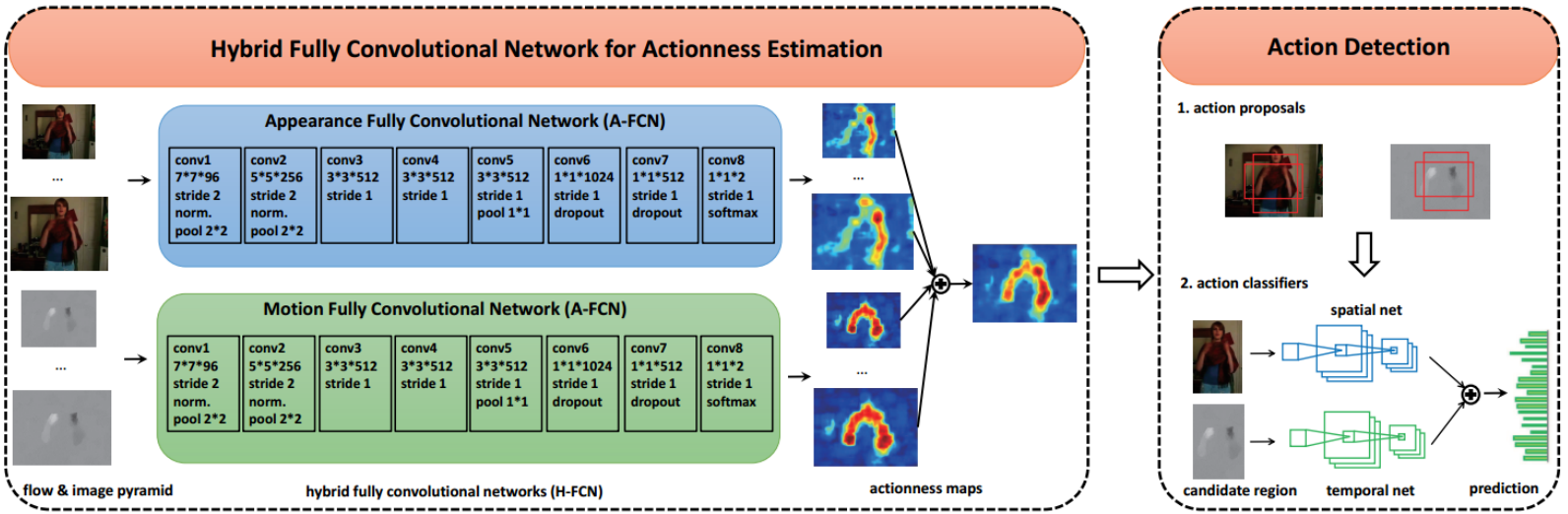


Figure 2. **Pipeline of our approach**. We propose a new architecture, called hybrid full convolutional network (H-FCN), for the task of actionness estimation. H-FCN contains two parts, namely appearance fully convolutional network (A-FCN) and motion fully convolutional network (F-FCN), which captures the visual cues from the perspectives of static appearance and dynamic motion, respectively. Based the estimated actionness maps, we design a RCNN-alike [10] action detection system, by first using actionness to generate action proposals and then applying two-stream convolutional networks to classify these proposals.

32

[32] Limin Wang et al. "Actionness estimation using hybrid fully convolutional networks". In: *CVPR*. 2016, pp. 2708–2717.

Singh et al. 2017 ICCV

Problems:

- **Real-time** multiple spatio-temporal (S/T) action localization and classification.

Limitations of past approaches:

- Offline approaches are time-consuming, hard to apply into real application, which acquires fast response.
- And the accuracy is not enough.

Motivation:

- Take advantage of more recent **SSD object detector** to address issues with **accuracy and speed at frame level**. (single-stage)

33

[33] Gurkirt Singh et al. "Online real-time multiple spatiotemporal action localisation and prediction". In: *ICCV*. 2017, pp. 3637–3646.

Singh et al. 2017 ICCV

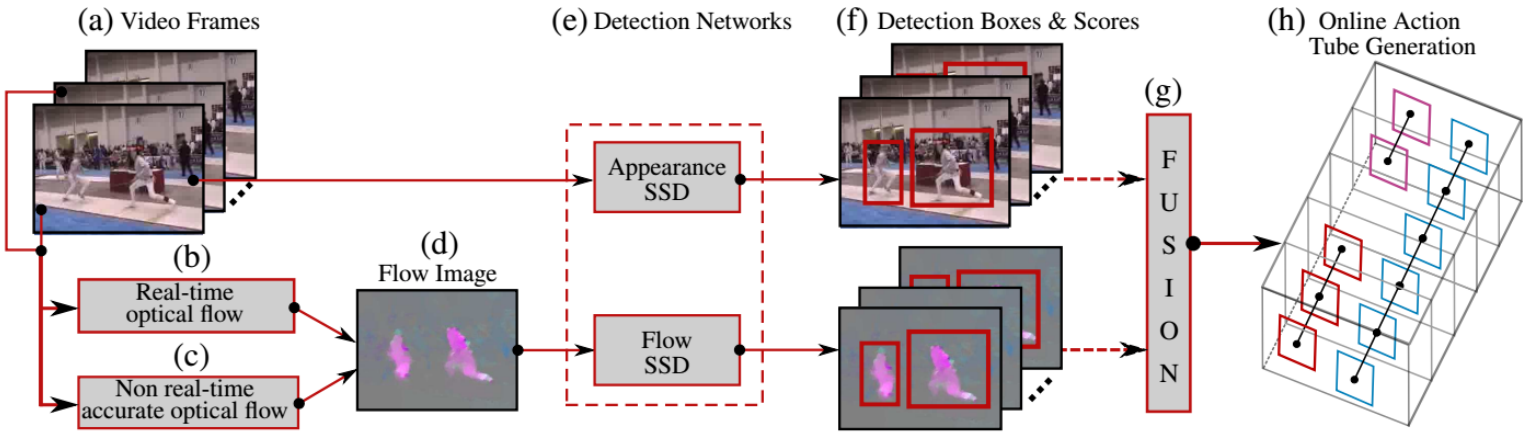


Figure 2. *At test time, the input to the framework is a sequence of RGB video frames* ***(a)****. A real-time optical flow (OF) algorithm* ***(b)*** *[16] takes the consecutive RGB frames as input to produce flow images* ***(d)****. As an option,* ***(c)*** *a more accurate optical flow algorithm [1] can be used (although not in real time).* ***(e)*** *RGB and OF images are fed to two separate SSD detection [22] networks (§ 3.2).* ***(f)*** *Each network outputs a set of detection boxes along with their class-specific confidence scores (§ 3.2).* ***(g)*** *Appearance and flow detections are fused (§ 3.3). Finally* ***(h)****, multiple action tubes are built up in an online fashion by associating current detections with partial tubes (§ 3.4).*

Singh et al. 2017 ICCV

Table 1. S/T action localisation results (mAP) on untrimmed videos of UCF101-24 dataset in split1.

IoU threshold δ	0.2	0.5	0.75	0.5:0.95
Yu *et al.* [58]‡	26.5	–	–	–
Weinzaepfel *et al.* [53]‡	46.8	–	–	–
Peng and Schmid [28]†	**73.5**	32.1	02.7	07.3
Saha *et al.* [33]†	66.6	36.4	07.9	14.4
Ours-Appearance (A)*	69.8	40.9	**15.5**	18.7
Ours-Real-time-flow (RTF)*	42.5	13.9	00.5	03.3
Ours-A + RTF (boost-fusion)*	69.7	41.9	14.1	18.4
Ours-A + RTF (union-set)*	70.2	43.0	14.5	19.2
Ours-Accurate - flow (AF)**	63.7	30.8	02.8	11.0
Ours-A + AF (boost-fusion)**	73.0	44.0	14.1	19.2
Ours-A + AF (union-set)**	**73.5**	**46.3**	15.0	**20.4**
SSD+ [33] A + AF (union-set)†	71.7	43.3	13.2	18.6

‡ These methods were using different annotations to [28, 33] and ours.
* Incremental & real-time ** Incremental, non real-time † Offline

Table 2. S/T Action localisation results (mAP) on J-HMDB-21.

IoU threshold δ	0.2	0.5	0.75	0.5:0.95
Gkioxari and Malik [7]†	–	53.3	–	–
Wang *et al.* [52]†	–	56.4	–	–
Weinzaepfel *et al.* [53]†	63.1	60.7	–	–
Saha *et al.* [33]†	72.6	71.5	43.3	40.0
Peng and Schmid [28]†	**74.1**	**73.1**	–	–
Ours-Appearance (A)*	60.8	59.7	37.5	33.9
Ours-Real-time-flow (RTF)*	56.9	47.4	20.2	19.3
Ours-A + RTF (union-set)*	66.0	63.9	35.1	34.4
Ours-A + RTF (boost-fusion)*	67.5	65.0	36.7	38.8
Ours-Accurate - flow (AF)**	68.5	67.0	38.7	36.1
Ours-A + AF (union-set)**	70.8	70.1	43.7	39.7
Ours-A + AF (boost-fusion)**	73.8	72.0	**44.5**	**41.6**
SSD+ [33] A + AF (boost-fusion)†	73.2	71.1	40.5	38.0

* Incremental & real-time ** Incremental, non real-time † Offline

Singh et al. 2017 ICCV

Comments:

- An extension of SSD to action detection at frame-level.
- Address the speed and accuracy.
- And the frame-level computation is quite suitable for the online action detection, and can be performed for early action prediction.
- But quite simple deep learning method, not to mention any essential problems in the field, such as temporal information, how to distinguish the temporal extent.
- And action detection at frame-level can be ambiguous.
- They do not exploit the temporal continuity of videos as they treat the video frames as a set of independent images on which a detector is applied independently.

Kalogeitone et al. 2017 ICCV

Motivation:

- Past CNN approaches do not use the temporal information, treating each frame individually, which may cause ambiguous.

Approaches:

- Use a sequence of frames as input to stack features for classification and 3D regression, based on anchor cuboid.
- Then link the cuboid to generate the tubes.

34

[34] Vicky Kalogeiton et al. “Action tubelet detector for spatio-temporal action localization”. In: *ICCV*. 2017, pp. 4405–4413.

Kalogeitone et al. 2017 ICCV

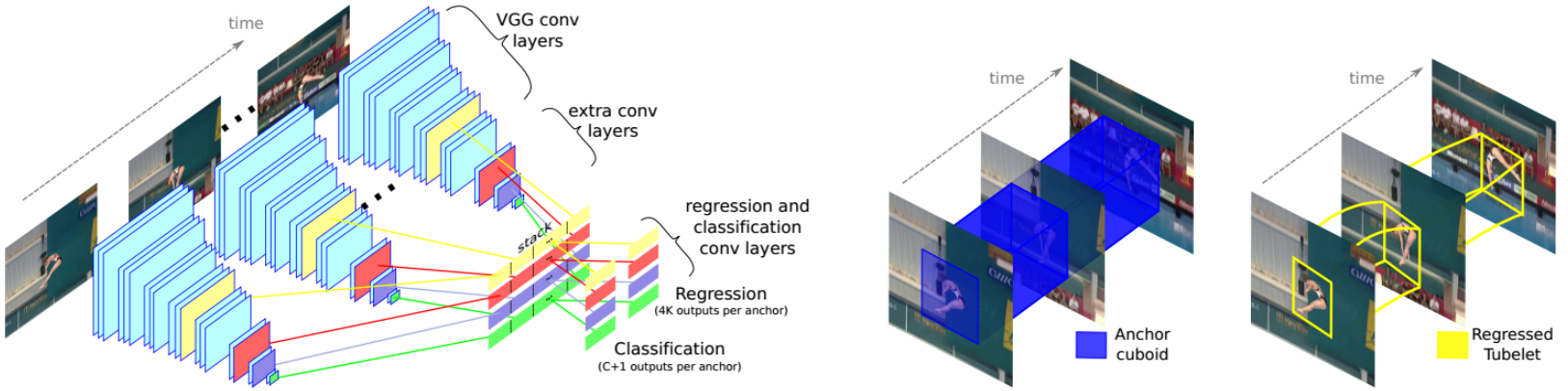


Figure 2. Overview of our ACT-detector. Given a sequence of frames, we extract convolutional features with weights shared between frames. We stack the features from subsequent frames to predict scores and regress coordinates for the anchor cuboids (middle figure, blue color). Depending on the size of the anchors, the features come from different convolutional layers (left figure, color coded: yellow, red, purple, green). As output, we obtain tubelets (right figure, yellow color).

Kalogeitone et al. 2017 ICCV

- A easy approach to perform action tubelets generation based on temporal information by stacked CNN features.
- One-stage method is faster.
- The result is still not obtained straightforward, needing to post-processing.
- And the information is not filtered, very rude.

Hou et al. 2017 ICCV

Motivation:

- Generalization from 2D to 3D does not take the temporal information into account and is not sufficiently expressive to distinguish between actions.

Approaches:

- Tube proposals: equal length clips is fed into Tube Proposal Network(TPN) based on 3D CNN, output a set of tube proposal. Then link them to form a complete tube proposals.
- Action label prediction: Tube-of-interest (ToI) pooling is applied to the linked action tube proposal to generate feature vector for action label prediction.

35

[35] Rui Hou, Chen Chen, and Mubarak Shah. "Tube convolutional neural network (T-CNN) for action detection in videos". In: *ICCV*. 2017, pp. 5822–5831.

Hou et al. 2017 ICCV

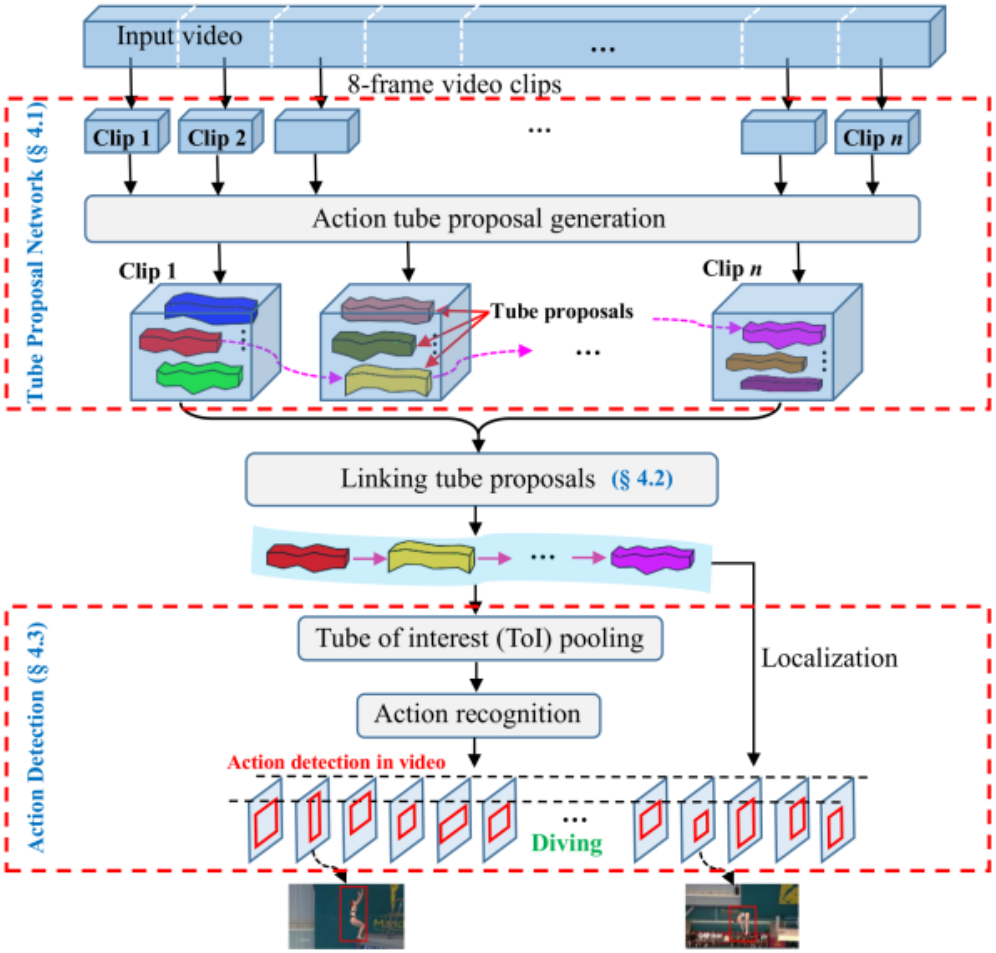


Figure 1: Overview of the proposed Tube Convolutional Neural Network (T-CNN).

Hou et al. 2017 ICCV

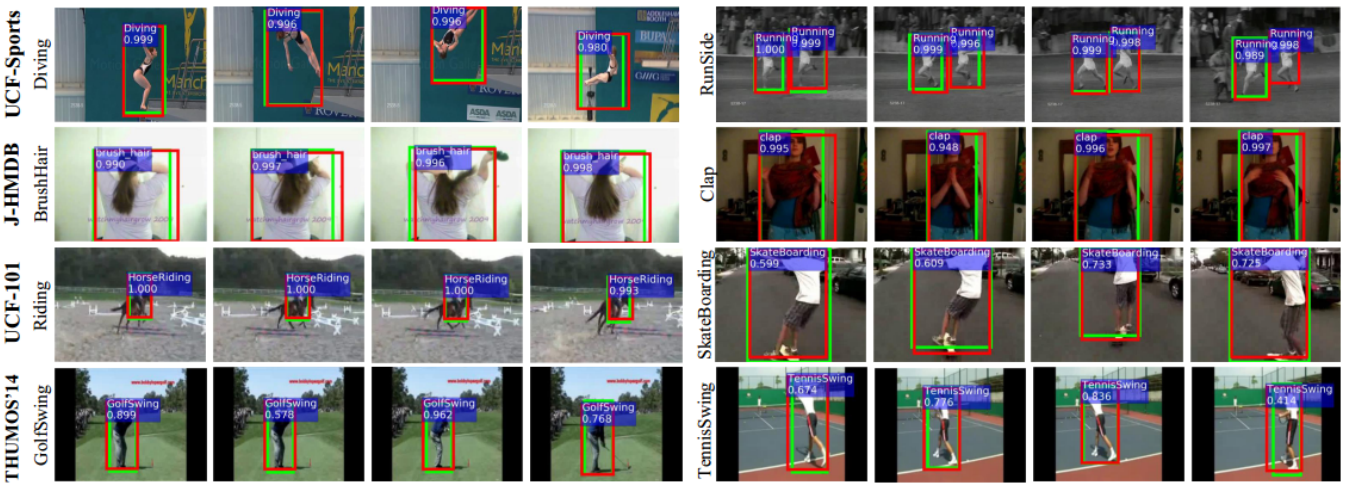


Figure 6: Action detection results by T-CNN on UCF-Sports, JHMDB, UCF-101 and THUMOS'14. Red boxes indicate the detections in the corresponding frames, and green boxes denote ground truth. The predicted label is overlaid.

Hou et al. 2017 ICCV

	f.-mAP ($\alpha = 0.5$)	v.-mAP ($\alpha = 0.2$)	v.-mAP ($\alpha = 0.5$)
Gkioxari *et al.* [5]	36.2	–	53.3
Weinzaepfel *et al.* [30]	45.8	63.1	60.7
Peng *et al.* [19]	58.5	74.3	73.1
Ours w/o skip pooling	47.9	66.9	58.6
Ours	**61.3**	**78.4**	**76.9**

Table 3: Comparison to the state-of-the-art on J-HMDB. The IoU threshold α for frame m-AP is fixed to 0.5.

	f.-mAP	video-mAP			
IoU th.		0.05	0.1	0.2	0.3
Weinzaepfel *et al.* [30]	35.84	54.3	**51.7**	46.8	37.8
Peng *et al.* [19]	39.63	54.5	50.4	42.3	32.7
Ours	**41.37**	**54.7**	51.3	**47.1**	**39.2**

Hou et al. 2017 ICCV

- A generlization of Peng R-CNN to 3D CNN, to combine temporal information.
- Still a two-stage method, and produce action tubes based on actionness score.
- Better performance than peng et al.

Saha et al. 2016 BMVC

Motivation:

- Apply R-CNN to action detection. (2017 online is their extension.)

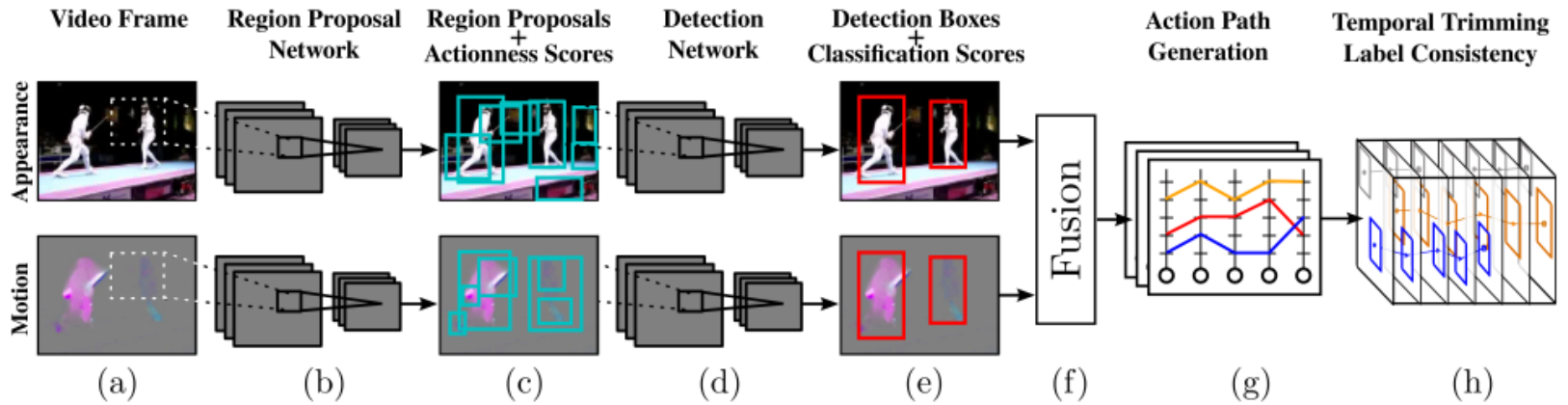


Figure 1: *At test time, **(a)** RGB and optical-flow images are passed to **(b)** two separate region proposal networks (RPNs). **(c)** Each network outputs region proposals with associated actionness scores (§ 3.1). **(d)** Each appearance/motion detection network takes as input the relevant image and RPN-generated region proposals, and **(e)** outputs detection boxes and softmax probability scores (§ 3.2). **(f)** Appearance and motion based detections are fused (§ 3.3) and **(g)** linked up to generate class-specific action paths spanning the whole video. **(h)** Finally the action paths are temporally trimmed to form action tubes (§ 3.4).*

36

[36] BMVC_framework.

Singh et al 2016 CVPR

Problems:

- **Fine-grained action detection**, where ”fine-grained” means that the differences among the classes of actions to be detected are small, like cooking scenarios.

Motivation:

- Apply LSTM to take advantage of temporal information.
- For static camera, use trajectory than optical flow is a better choices.

37

[37] Bharat Singh et al. “A multi-stream bi-directional recurrent neural network for fine-grained action detection”. In: *CVPR*. 2016, pp. 1961–1970.

Singh et al 2016 CVPR

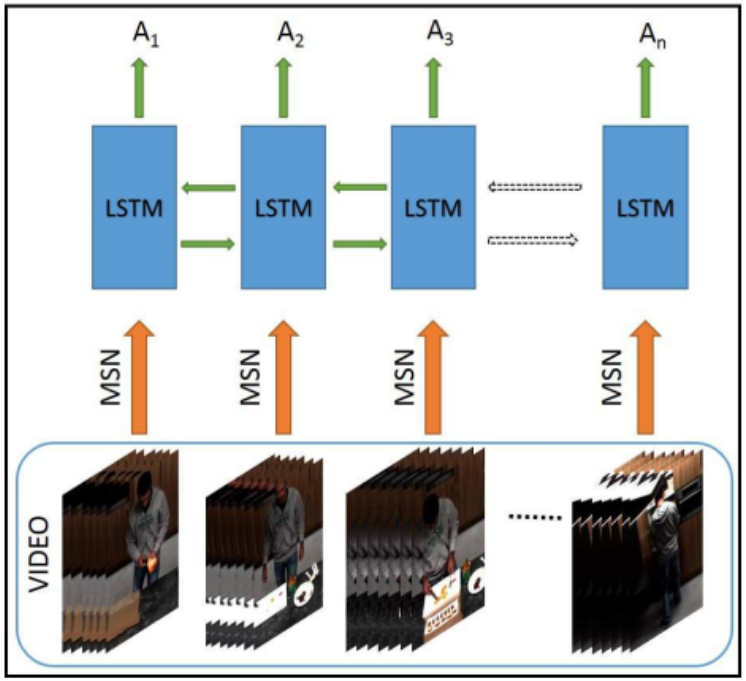


Figure 1. Framework for our approach. Short chunks of a video are given to a multi-stream network (MSN) to create a representation for each chunk. The sequence of these representations is then given to a bi-directional LSTM, which is used to predict the action label, A_i. Details of the multi-stream network are shown in Fig. 2.

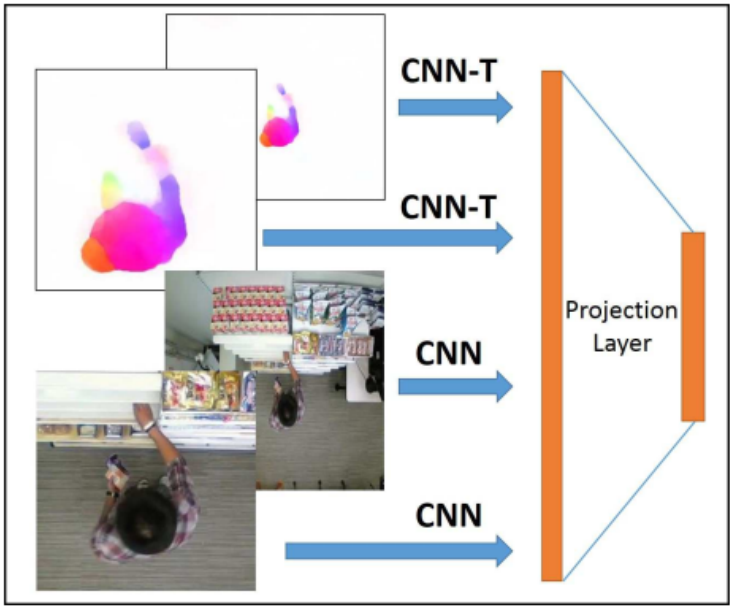


Figure 2. Figure depicting our multi-stream network (MSN). The multi-stream network uses two different streams of information (motion and appearance) for each of two different spatial croppings (full-frame and person-centric) to analyze short chunks of video. One network (CNN-T) computes features on pixel trajectories (motion), while the other (CNN) computes features on RGB channels (appearance).

Singh et al 2016 CVPR

Comments:

- What's the difference between fine-grained and general action detection?
- Camera static? Little inter-class difference? More hard on localization or classification? More need temporal information? More need to detect interaction between objects?
- Yes, **more interaction or relationship between objects**.

Mettes et al. 2017 ICCV

Problems:

- Localization and classification of human actions **without** the need for any video training examples.

Motivation:

- Address the **interaction** between human and objects, which is overlooked by past zero-shot approaches, which only encode semantic information of objects and attributes.

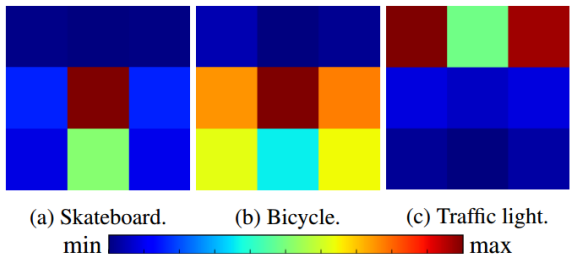


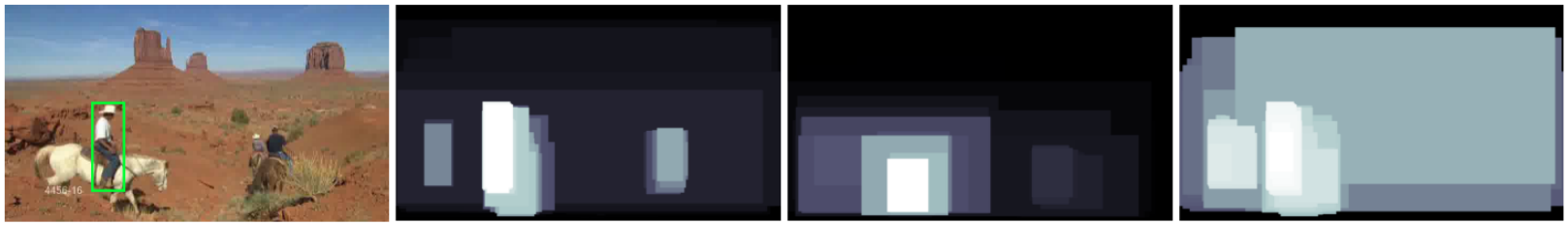
Figure 2: **Examples of preferred spatial relations of objects relative to actors.** In line with our intuition, skateboards are typically on or below the actor, while bicycles are typically to the left or right of actors and traffic lights are above the actors.

[38] Pascal Mettes and Cees GM Snoek. "Spatial-aware object embeddings for zero-shot localization and classification of actions". In: *ICCV*.

Mettes et al. 2017 ICCV

Approach:

- Gather prior knowledge on actions, actors, objects, and their interactions;
- Compute spatial-aware embedding scores for bounding boxes;
- Link boxes into action tube.



(a) Video frame. (b) Actor detection. (c) Object detection (horse). (d) Spatial relation match.

Figure 3: **Example of our spatial-aware embedding.** The actor sitting on the left horse (green box) is most relevant for the action *Riding horse* based on the actor detection, horse detection, and spatial relations between actors and horses.

Mettes et al. 2017 ICCV

	Localization (mAP @ 0.5)				**Classification** (mean accuracy)			
	# local objects				# local objects			
	0	1	2	5	0	1	2	5
Embedding I: ***Actor-only***	0.083	-	-	-	0.100	-	-	-
Embedding II: ***Actors and objects***	-	0.175	0.182	0.193	-	0.205	0.117	0.139
Embedding III: ***Spatial-aware***	-	**0.221**	0.209	0.199	-	0.180	0.196	**0.255**

Table 1: **Influence of spatial awareness.** On UCF Sports we compare our spatial-aware object embedding to two other embeddings; using only the actors and using actors with objects, while ignoring their spatial relations. Our spatial-aware embedding is preferred for both localization (one object per action) and classification (five objects per action).

Mettes et al. 2017 ICCV

Comments:

- Address the interaction for the action detection.
- Maybe it's the next direction for this task.

Li et al. 2018 ECCV

Motivation:

- Action recognition often requires reasoning about the actor's **relationship** with objects and other actors, both spatially and temporally.
- Extract **actor-centric** relationships.

Problem:

- Action detection that can infer **actor-object spatio-temporal relations** automatically with only **actor-level supervision**.

39

[39] Dong Li et al. "Recurrent tubelet proposal and recognition networks for action detection". In: *ECCV*. 2018, pp. 303–318.

Li et al. 2018 ECCV

Approach:

- Inspired by VQA Santoro, build a pairwise relationship module to produce a feature that represents the relationship, which is beneficial for the classification and regression.

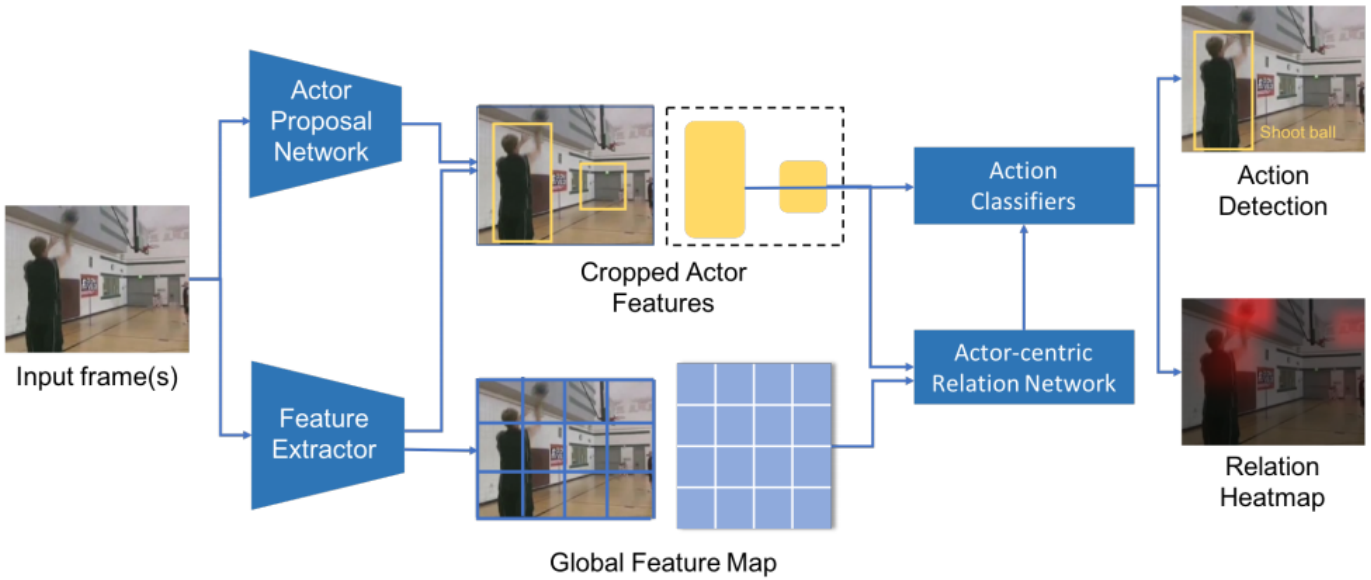


Fig. 2: Overview of our proposed action detection framework. Compared to a standard action detection approach, the proposed framework extracts pairwise relations from cropped actor features and a global feature map with the actor-centric relation network (ACRN) module. These relation features are then used for action classification.

Li et al. 2018 ECCV

Model	frame-AP	video-AP
Peng et al. [37]	58.5	73.1
ACT [24]	65.7	73.7
I3D [13]	73.3	78.6
Base-Model	75.2	78.8
ACRN	**77.9**	**80.1**

(a) JHMDB (3 splits)

Model	frame-AP
Single frame [13]	14.2
I3D [13]	15.1
Base-Model	15.5
ACRN	**17.4**

(b) AVA (version 2.1)

Table 3: Comparison with state of the art on (a) the JHMDB dataset and (b) AVA . For JHMDB, we report average precision over 3 splits.

Li et al. 2018 ECCV

Comments:

- A implicit expression of the relationship between actor and objects.
- Inspired by the relationship built by AVA dataset, the attributes of object and background is also essential for the recognition of action, besides the actor themselves.
- But the type of relationship is also need to discuss? The position or just the attributes?

Sun et al. 2018 ECCV

Motivation:

- Existing approaches predominately generate action proposals for each individual frame or fixed-length clip independently, while overlooking temporal context across them.

Approach:

- Recurrent Tubelet Proposal(RTP) network: Given the features of I(t) and I(t+1) and the proposal b(i), estimate the actionness score and the movement of the proposal m(i+1), where b(i+1) = b(i) + m(i+1). The proposal b(i) is used to extract the features of corresponding areas.
- Recurrent Tubelet Recognition(RTR) network: A LSTM network, feded by three features: proposal-cropped feature, RoI pooling, whole images.

[40]

[40] Chen Sun et al. "Actor-centric relation network". In: *ECCV*. 2018, pp. 318–334.

Sun et al. 2018 ECCV

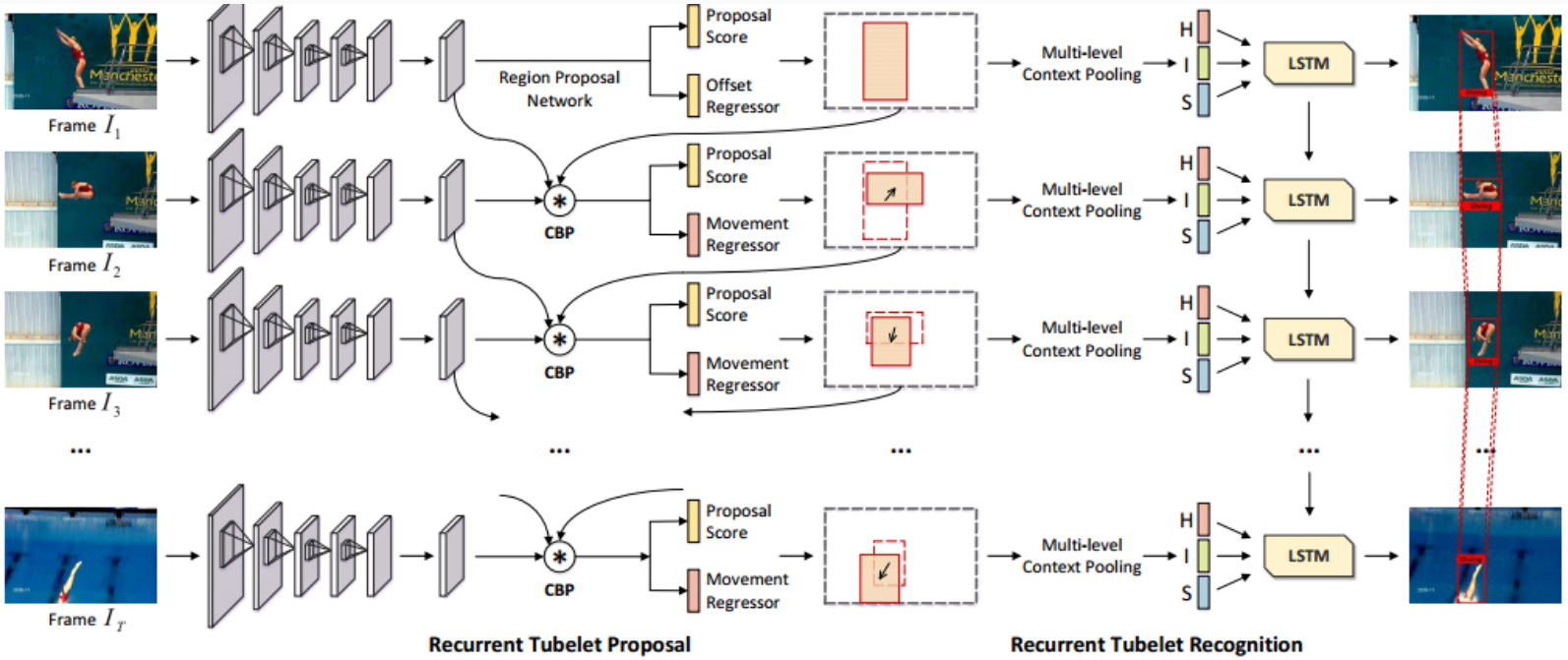


Fig. 2. The overview of RTPR networks. It consists of two components: RTP networks and RTR networks. CNNs are first utilized in RTP for feature extraction. Then RPN is applied on the start frame for proposal initialization. RTP estimates the movements of proposals in current frame to produce proposals in next frame. After proposal linking and temporal trimming, obtained tubelet proposals are fed into RTR. The RTR employs a multi-channel architecture to capture different semantic-level information, where an individual LSTM is utilized to model temporal dynamics on each channel.

Sun et al. 2018 ECCV

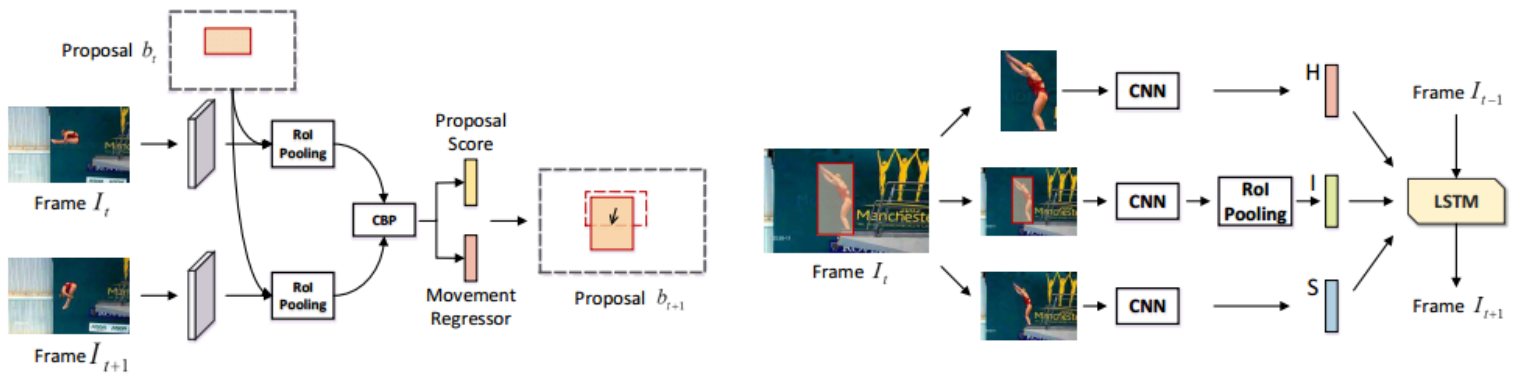


Fig. 3. (a) RTP networks. Two consecutive frames I_t and I_{t+1} are first fed into CNNs. Given the proposal b_t in current frame I_t, we perform RoI pooling on both I_t and I_{t+1} w.r.t. the same proposal b_t. The two pooled features are fed into a CBP layer to generate the correlation features, which are used to estimate the movement of proposal b_t and the actionness score. **(b)** RTR networks. We capitalize on a multi-channel network for tubelet recognition. Three different semantic clues, i.e., *human only* (H), *human-object interaction* (I), and *scene* (S), are exploited, where the features on proposal-cropped image, the features with RoI pooling on the proposal, and the features on whole frame are extracted. Each of them is fed into an LSTM to model the temporal dynamics.

Sun et al. 2018 ECCV

Method	RGB	Flow	Fusion
[8] w/ ResNet-101	17.1	9.3	18.4
RTPR w/ VGG-16	19.4	15.6	20.1
RTPR w/ ResNet-101	**20.5**	**16.2**	**22.3**

Fig. 7. Four detection examples of our method from UCF-Sports, J-HMDB, UCF-101, and AVA. The proposal score is given for each bounding box. Top predicted action classes for each tubelet are on the right. Red labels indicate ground-truth.

Table 2. Performance contribution of each component in the proposed RTPR. U-S, J-H, and U-1 represent UCF-Sports, J-HMDB (split 1), and UCF-101 respectively.

Method	RTP	LSTM	HIS	Flow	U-S	J-H	U-1	AVA
Faster R-CNN					83.8	56.5	56.0	15.6
+RTP	√				85.2	58.2	57.9	16.8
+LSTM	√	√			85.7	59.1	59.2	18.2
+HIS	√	√	√		87.6	61.7	60.4	19.4
RTPR	√	√	√	√	**97.8**	**80.7**	**76.3**	**20.1**

Table 3. Video-mAP comparisons on UCF-Sports, J-HMDB, and UCF-101.

Method	UCF-Sports		J-HMDB					UCF-101			
	0.2	0.5	0.1	0.2	0.3	0.4	0.5	0.05	0.1	0.2	0.3
Gkioxari *et al.* [7]	-	75.8	-	-	-	-	53.3	-	-	-	-
Weinzaepfe *et al.* [38]	-	90.5	-	63.1	63.5	62.2	60.7	54.3	51.7	46.8	37.8
Saha *et al.* [28]	-	-	72.7	72.6	72.6	72.2	71.5	79.1	76.6	66.8	55.5
Peng *et al.* [22][1]	94.8	94.7	-	74.3	-	-	73.1	78.8	77.3	72.9	65.7
Singh *et al.* [31]	-	-	-	73.8	-	-	72.0	-	-	73.5	-
Kalogeiton *et al.* [14]	92.7	92.7	-	74.2	-	-	73.7	-	-	77.2	-
Hou *et al.* [11][2]	95.2	95.2	-	78.4	-	-	76.9	78.2	77.9	73.1	69.4
Yang *et al.* [39]	-	-	-	-	-	-	-	79.0	77.3	73.5	60.8
He *et al.* [9]	96.0	95.7	79.8	79.7	79.3	78.5	77.0	-	-	71.7	-
RTPR											
-w/ VGG-16	97.8	97.8	**83.0**	82.3	82.0	81.2	80.5	81.5	80.7	76.3	70.9
-w/ ResNet-101	**98.6**	**98.6**	**83.0**	**82.7**	**82.3**	**82.3**	**81.3**	**82.1**	**81.3**	**77.9**	**71.4**

Sun et al. 2018 ECCV

Comments:

- A LSTM-based model to incorporate the temporal information.
- Multi-channel architecture to incorporate the proposal of previous frame to enhance the recognition.
- Straightforward network to frame-by-frame predict the result.

Duarte et al. 2018 NIPS

Motivation:

- Generalize capsule network from 2D to 3D for action detection.

Table 1: Action localization accuracy of VideoCapsuleNet. The results reported in the row VideoCapsuleNet* use the ground-truth labels when generating the localization maps, so they should not be directly compared with the other state-of-the-art results.

Method	UCF-Sports		J-HMDB		UCF-101				
	f-mAP	v-mAP	f-mAP	v-mAP	f-mAP	v-mAP			
	0.5	0.2	0.5	0.2	0.5	0.1	0.2	0.3	0.5
Saha et al. Saha *et al.* (2016)	-	-	-	72.6	-	76.6	66.8	55.5	35.9
Peng et al. Peng & Schmid (2016)	84.5	94.8	58.5	74.3	65.7	77.3	72.9	65.7	35.9
Singh et al. Singh *et al.* (2017)	-	-	-	73.8	-	-	73.5	-	46.3
Kalogeiton et al. Kalogeiton *et al.* (2017)	87.7	92.7	65.7	74.2	69.5	-	77.2	-	51.4
Hou et al. Hou *et al.* (2017)	86.7	95.2	61.3	78.4	67.3	77.9	73.1	69.4	-
Gu et al. Gu *et al.* (2018)	-	-	73.3	-	76.3	-	-	-	59.9
He et al. He *et al.* (2018)	-	96.0	-	79.7	-	-	71.7	-	-
VideoCapsuleNet	83.9	**97.1**	64.6	**95.1**	**78.6**	**98.6**	**97.1**	**93.7**	**80.3**
VideoCapsuleNet*	82.8	97.1	66.8	95.4	80.1	98.9	97.4	94.2	82.0

Metrics We compute frame-mAP and video-mAP for the evaluation Peng & Schmid (2016). For frame-mAP we set the IoU threshold at $\alpha = 0.5$, and compute the average precision over all the frames for each class. This is then averaged to obtain the f-mAP. For video-mAP the average precision is computed for the 3D IoUs at different thresholds over all the videos for each class, and then averaged to obtain the v-mAP.

41

[41] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. "Videocapsulenet: A simplified network for action detection". In: *NIPS*. 2018, pp. 7610–7619.

Yang et al. 2019 CVPR

Motivation:

- Two-stage process's tubelet building suffers from **spatial displacement** of action, and **temporal modeling**.
- Previous approaches either **assume small displacement of actions**, or **link the location offline**.

Approach:

- A multi-step optimization process that progressively refines the initial proposals towards the final solution.
- Address spatio-temporal localization, the process take a progressive manner to refine the location, through spatial refinement and temporal refinement.

42

[42] Xitong Yang et al. "Step: Spatio-temporal progressive learning for video action detection". In: *CVPR*. 2019, pp. 264–272.

Yang et al. 2019 CVPR

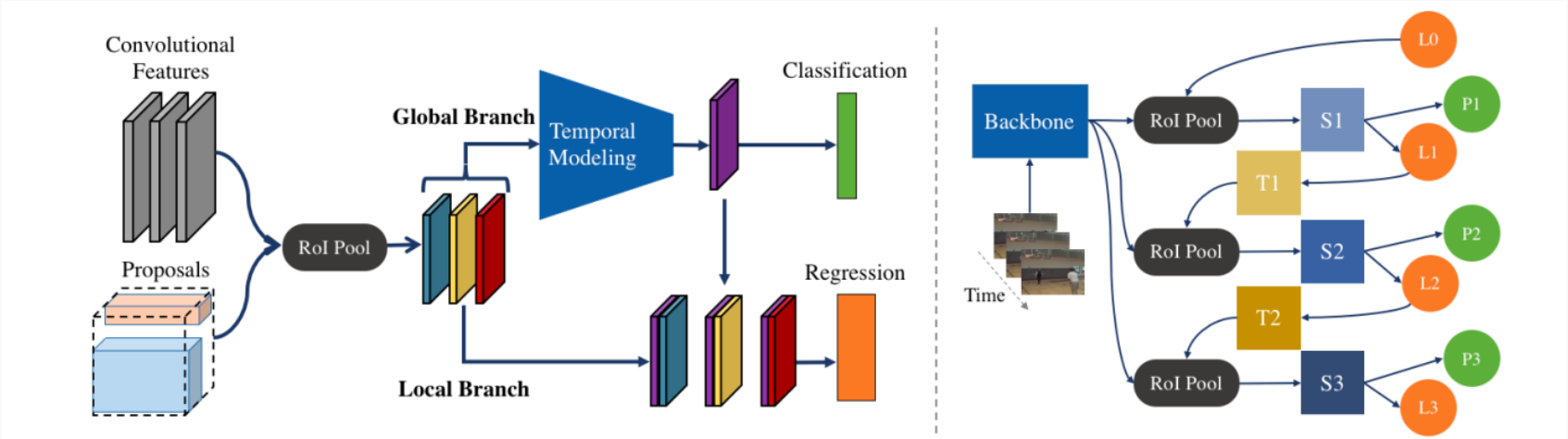


Figure 3: Left: the architecture of our two-branch network. Right: the illustration of our progressive learning framework, where "S" indicates spatial refinement, "T" temporal extension, "P" classification, and "L" localization, the numbers correspond to the steps, and "L0" denotes the initial proposals.

Yang et al. 2019 CVPR

Method	frame-mAP	video-mAP		
	0.5	0.05	0.1	0.2
MR-TS [26]	65.7	78.8	77.3	72.9
ROAD [30]	-	-	-	73.5
CPLA [37]	-	79.0	77.3	73.5
RTPR [20]	-	81.5	80.7	76.3
PntMatch [38]	67.0	79.4	77.7	76.2
T-CNN [16]	67.3	78.2	77.9	73.1
ACT [17]	69.5	-	-	76.5
Ours	**75.0**	**84.6**	**83.1**	**76.6**

Table 2: Comparison with the state-of-the-art methods on UCF101 by frame-mAP (%) and video-mAP (%) under different IoU thresholds.

Method	frame-mAP
Single Frame* [12]	14.2
I3D [12]	14.7
I3D* [12]	15.6
ACRN* [32]	17.4
Ours	**18.6**

Table 3: Comparison with the state-of-the-art methods on AVA by frame-mAP (%) under IoU = 0.5. "*" means the

Yang et al. 2019 CVPR

Comments:

- **Break the fixed two-stage process for action detection, treat the regression and classification as side-by-side tasks.**
- Personally, I also think the two-stage method is wired (what's the actionness), if you don't know the action class, how can you get your action tube, and how can you define the start and the end of the action. The only you know is object detection in videos. Maybe several keyframe action classification is enough. But the localization, especially the temporal extent distinguish is hard.
- I understand why this paper is oral, while others are not.

Zhang et al. 2019 CVPR

Motivation:

- Incorporate domain knowledge into the structure of the model to simplify the action detection problem.
- Inspired by only several cues are crucial for action detection: location, motion, interactions.
- Poor temporal detection in AVA.

Approach:

- Explicit model long-term human behaviour and human-human and human-object interaction.
- 1. Video detect the actors and objects;
- 2. Track to generate tubes;
- 3. Construct an actor-centric graph;
- 4. Aggregate features to predict the action.

[43]

[43] Yubo Zhang et al. "A structured model for action detection". In: *CVPR*. 2019, pp. 9975–9984.

Zhang et al. 2019 CVPR

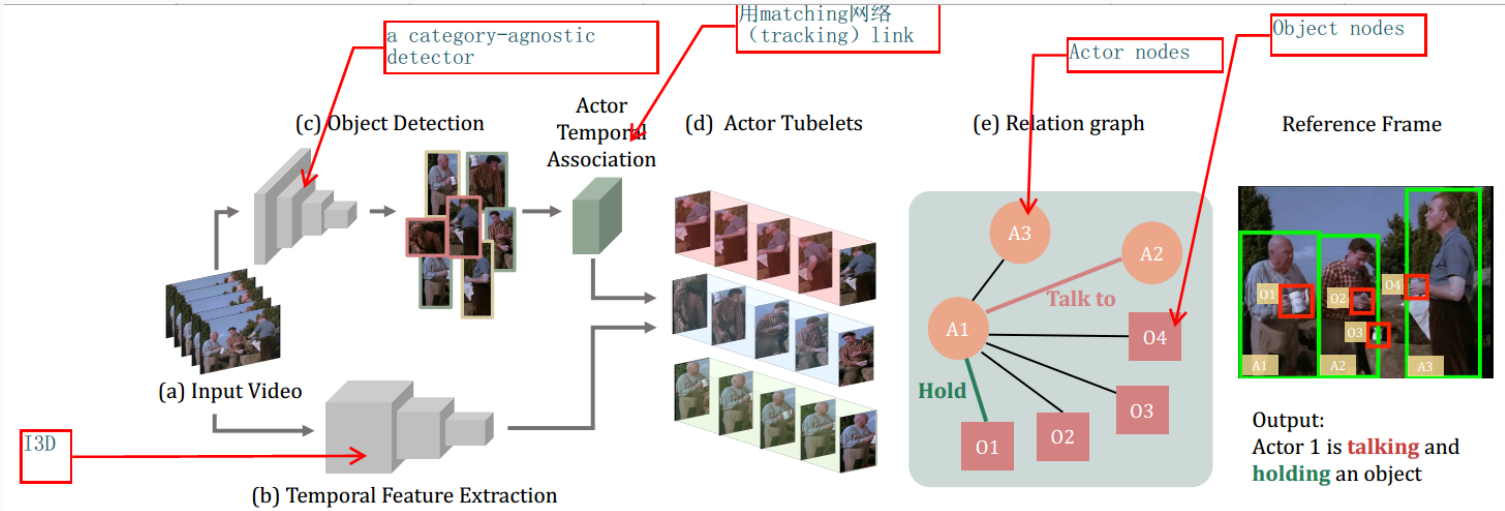


Figure 2. Overview of our proposed framework. We model both long-term person behaviour and human-human, human-object interactions structurally in a unified framework. The actors across the video are associated to generate actor tubelets for learning long temporal dependency. The features from actor tubelets and object proposals are then used to construct a relation graph to model human-object manipulation and human-human interaction actions. The output of our model are actor-centric actions.

Model	mAP
Single Frame model [15]	14.2
ACRN [50]	17.4
Our model	22.2

Table 3. Comparison of our model to the state-of-the-art methods on the validation set of AVA.

Zhang et al. 2019 CVPR

Comments:

- Inspired by the interaction built by AVA, the approach further explicitly models the relationships.
- In fact, combining the pairwise interaction information into network will enhance the performance.
- But we don't know the bottleneck is the classification or the regression of spatial or temporal?

Zhao et al. 2019 CVPR

Motivation:

- The two-stream framework cannot produce satisfactory results, even though double computation.

Approach:

- Use motion as a attention condition, fuse the two features into one spatio-temporal features.

44

[44] Yubo Zhang et al. “A structured model for action detection”. In: *CVPR*. 2019, pp. 9975–9984.

Zhao et al. 2019 CVPR

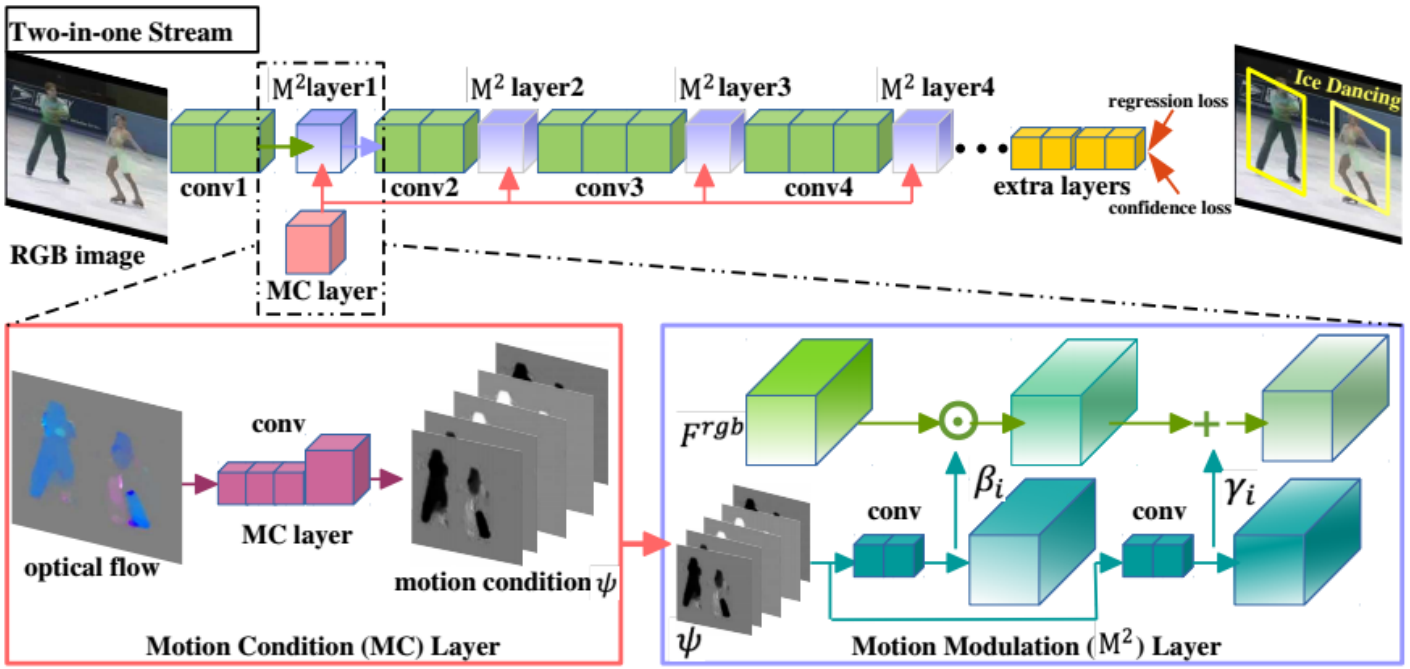


Figure 2: **Two-in-one network architecture**. The motion condition layer (pink cube) maps flow images to prior condition information. The condition inputs to the motion modulation layer (purple cube) to generate transformation parameters which are used to modulate RGB features (F^{rgb}). The network has half the computation and parameters of a two-stream equivalent, while obtaining better action detection accuracy.

Zhao et al. 2019 CVPR

	UCF101-24				*UCFSports*			*J-HMDB*		
	0.20	0.50	0.75	0.50:0.95	0.50	0.75	0.50:0.95	0.50	0.75	0.50:0.95
Single-frame										
Peng & Schmid [29]	71.80	35.90	1.60	8.80	**94.80**	47.30	51.00	70.60	*48.20*	*42.20*
Saha *et al.* [34]	66.70	35.90	7.90	14.40	–	–	–	*71.50*	43.30	40.00
Behl *et al.* [1]	71.53	40.07	13.91	17.90	–	–	–	–	–	–
Singh *et al.* [37]	73.50	46.30	15.00	20.40	–	–	–	**72.00**	44.50	41.60
This paper: Two-in-one	*75.13*	*47.47*	*17.21*	*21.51*	87.46	*57.81*	*51.69*	60.99	47.23	38.72
This paper: Two-in-one two stream	**77.49**	**49.54**	**17.62**	**22.02**	*87.81*	**62.67**	**52.32**	70.00	**52.00**	**43.20**
Multi-frame										
Saha *et al.* [33]	63.06	33.06	0.52	10.72	–	–	–	57.31	–	–
Kalogeiton *et al.* [21]	76.50	49.20	19.70	23.40	92.70	78.40	58.80	*73.70*	*52.10*	*44.80*
Singh *et al.* [36]	**79.00**	**50.90**	20.10	*23.90*	–	–	–	–	–	–
This paper: Two-in-one	75.48	48.31	*22.12*	*23.90*	*92.74*	*83.64*	*59.60*	57.96	42.78	34.56
This paper: Two-in-one two stream	*78.48*	*50.30*	**22.18**	**24.47**	**96.52**	**90.41**	**63.59**	**74.74**	**53.28**	**45.01**

Table 3: **Accuracy comparison to the state-of-the-art.** Bold means top accuracy and italic means second top accuracy. For the high overlap setting of $mAP@IoU$=0.5:0.95, our two-in-one stream works well in both a single-frame and multiple-frame network for all three datasets. When we add an additional flow-stream to obtain a two-in-one two stream we further improve accuracy.

Song et al. 2019 CVPR

Motivation:

- Address the unsatisfactory result for temporal extent detection.

Approach:

- Additional modeling transition to enhance the capability of distinguished the action extent;
- Conv-LSTM unit as baseline;

45

[45] Jiaojiao Zhao and Cees GM Snoek. "Dance with Flow: Two-in-One Stream Action Detection". In: *CVPR*. 2019, pp. 9935–9944.

Song et al. 2019 CVPR

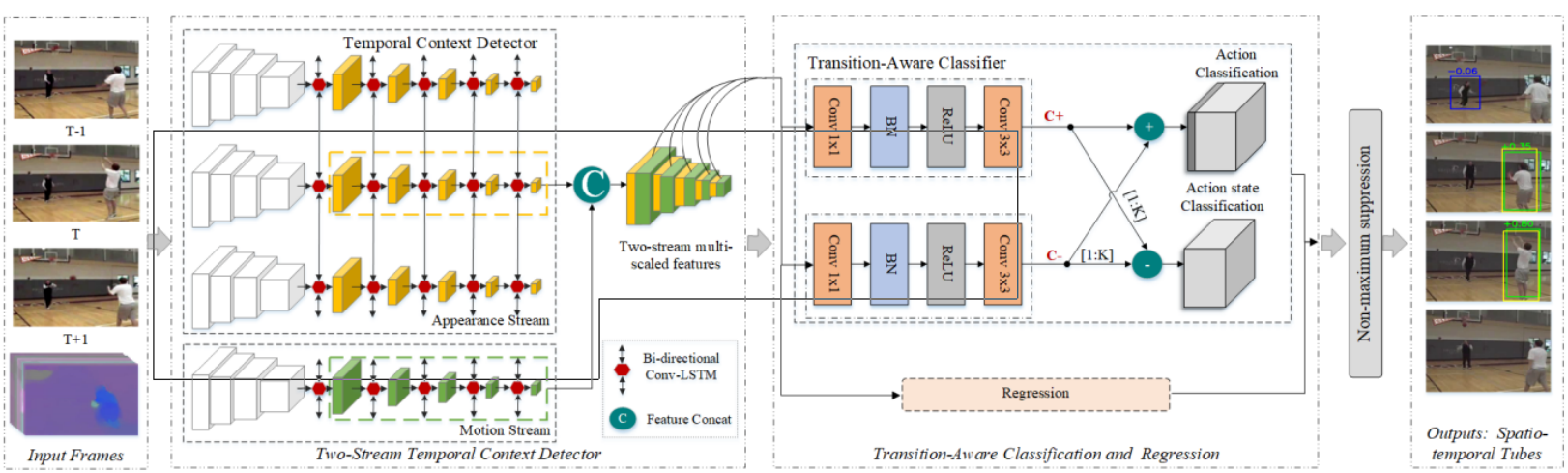


Figure 2. Overall framework of the proposed TACNet. TACNet mainly contains two modules: temporal context detector and transition-aware classifier. In the temporal context detector, we embed several multi-scale Conv-LSTM [11] units in the standard SSD detector [12] to extract temporal context. Based on the recurrent action detector, the transition-aware classifier is designed to simultaneously detect the action categories and states. Then, we can correctly localize the temporal boundaries for the target actions.

Song et al. 2019 CVPR

Table 3. Comparison with the state-of-the-art on J-HMDB (trimmed) and UCF101 (untrimmed)

Method	J-HMDB					UCF101-24 (Full)[1]					UCF101-24 (Untrimmed)[1]	
	F-mAP	Video-mAP				F-mAP	Video-mAP				F-mAP	Video-mAP
		0.2	0.5	0.75	0.5:0.95		0.2	0.5	0.75	0.5:0.95		0.5
Saha [17]	-	72.6	71.5	43.3	40.04	-	66.7	35.9	7.9	14.4	-	-
Peng [15]	58.5	74.3	73.1	-	-	65.7	73.5	32.1	2.7	7.3	-	-
Singh [19]	-	73.8	72.0	44.5	41.6	-	73.5	46.3	15.0	20.4	-	-
Hou [8]	61.3	**78.4**	**76.9**	-	-	41.4	47.1	-	-	-	-	-
Becattini [1]	-	-	-	-	-	-	67.0	35.7	-	-	-	-
Kalogeiton [10]	**65.7**	74.2	73.7	52.1	44.8	69.5	76.5	49.2	19.7	23.4	52.1	23.5
Ours	65.5	74.1	73.4	**52.5**	**44.8**	**72.1**	**77.5**	**52.9**	**21.8**	**24.1**	**58.0**	**31.3**

[1] UCF101-24 is a mixture dataset which is made up of untrimmed categories and trimmed categories, thus we evaluate our approaches in two criterions to fully illustrate the performance gain on untrimmed videos.

Song et al. 2019 CVPR

Comments:

- Address the temporal context problems, explicitly classification of them, for that the individual frame is hard to distinguish.
- Quite important for untrimmed videos.
- Maybe a bottleneck is the temporal context.

Mac et al. 2019 ICCV

Problems:

-

46

[46] Khoi-Nguyen C Mac et al. “Learning Motion in Feature Space: Locally-Consistent Deformable Convolution Networks for Fine-Grained Action Detection”. In: *ICCV*. 2019, pp. 6282–6291.

Action Detection

History:

- **Template-based**: describe the action as the multiple shapes with specific order in timeline. Can be formulated as a search problem, which can be solved by energy minimization or discriminative function.
- The shape cannot discriminate the difference of action.
- **Statistic-feature**-based: use bags-of-feature to describe the actions, then search in the 3D spatiotemporal space.
- Focus on the several parts in spatial or temporal space (human-centric, deformable).
- Other formulation: **generative model**(HMM) logics; **Regression problem**(map between video and position).

Action Detection

characteristics:

- Highly similar to object detection.
- Can be formulated as a search, regression or classification problem.

Challenges:

- Large intra-variance of action.
- Very large search space.

Questions:

- How to describe actions?
- How to search actions in such a large space?

Accuracy:

- Spatial location;
- Temporal location;
- Classification.

Categories

- Segmentation-based
- Motion-based
- Matching-based
- Discriminative etc.
- Features
- Template action detection.