Predicting the *Where* and *What* of actors and actions through Online Action Localization

Khurram Soomro, Haroon Idrees, Mubarak Shah
Center for Research in Computer Vision (CRCV), University of Central Florida (UCF)
{ksoomro, haroon, shah}@cs.ucf.edu

Abstract

This paper proposes a novel approach to tackle the challenging problem of 'online action localization' which entails predicting actions and their locations as they happen in a video. Typically, action localization or recognition is performed in an offline manner where all the frames in the video are processed together and action labels are not predicted for the future. This disallows timely localization of actions - an important consideration for surveillance tasks.

In our approach, given a batch of frames from the immediate past in a video, we estimate pose and oversegment the current frame into superpixels. Next, we discriminatively train an actor foreground model on the superpixels using the pose bounding boxes. A Conditional Random Field with superpixels as nodes, and edges connecting spatio-temporal neighbors is used to obtain action segments. The action confidence is predicted using dynamic programming on SVM scores obtained on short segments of the video, thereby capturing sequential information of the actions. The issue of visual drift is handled by updating the appearance model and pose refinement in an online manner. Lastly, we introduce a new measure to quantify the performance of action prediction (i.e. online action localization), which analyzes how the prediction accuracy varies as a function of observed portion of the video. Our experiments suggest that despite using only a few frames to localize actions at each time instant, we are able to predict the action and obtain competitive results to state-of-the-art offline methods.

1. Introduction

Predicting *what* and *where* an action will occur is an important and challenging computer vision problem for automatic video analysis [17, 35, 41, 4]. In many applications associated with monitoring and security, it is crucial to detect and localize actions in a timely fashion. A particular example is detection and localization of undesirable or malicious actions. There have been recent efforts to predict activities by early recognition [18, 16, 26, 14]. These methods only attempt to predict the label of the action, *what* of an action, without any localization. Thus, the important question about *where* an action is being performed cannot be answered easily.

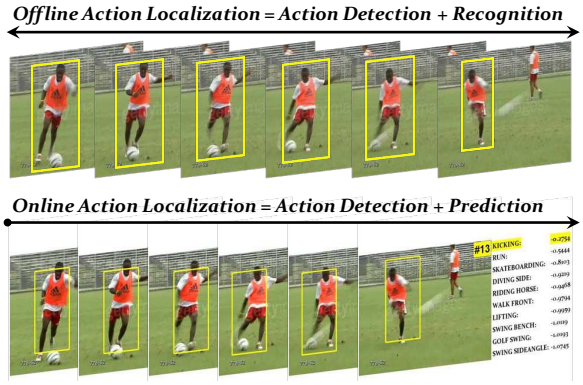


Figure 1. This figure illustrates the problem we address in this paper. The top row shows the case when we have an entire video to detect and recognize actions, i.e., *offline* action localization. The bottom row is an example of *online* action localization, which involves predicting the action class (e.g. *Kicking*) as well as the location of the actor in every frame, as the video is streamed.

Existing action localization methods [17, 35, 41, 29, 10, 27] classify and localize actions after completely observing an entire video sequence (top row in Fig. 1). The goal is to localize an action by finding the volume that encompasses an entire action. Some approaches are based on sliding-windows [29, 22], while others segment the video into supervoxels which are merged into action proposals [10, 21, 27]. The action proposals from either methods are then labeled using a classifier. Essentially, an action segment is classified after it has been

localized. Since offline methods have whole video at their disposal, they can take advantage of observing entire motion of action instances. In this paper, we address the problem of *Online Action Localization*, which aims at localizing an action and predicting its class label in a streaming video (see bottom row in Fig. 1). Online action localization involves the use of limited motion information in partially observed videos for frame-by-frame action localization and label prediction.

Low-level motion features, both hand-crafted [34] and deep learned [36] have imparted significant gains to the performance of action recognition and localization algorithms. They have been extensively employed in various action recognition methods. However, human actions inherently consists of articulation which low-level features cannot model explicitly. On the other hand, the compact and low-dimensional nature of high-level representations such as human poses (locations of different joints), makes them sensitive and unstable for action recognition. An incorrect estimation of pose translates to large variation in descriptors that aim to capture the configuration of joints both in space and time. This drawback can hamper the performance of any action localization and recognition algorithm. Nonetheless, a few methods (e.g. [35]) have successfully employed pose features for offline action localization.

In this paper, we propose to use the high level structural information using pose in conjunction with a superpixel based discriminative actor foreground model that distinguishes the foreground action and the background. This superpixel-based model incorporates visual appearance using color features, as well as structural cues through joint locations. Using the superpixel-based actor foreground model we generate a confidence map, that is later used to predict and locate the action segments by inferring on a Conditional Random Field. Since the appearance of an actor changes due to articulation and camera motion, we retrain foreground model as well as impose spatio-temporal constraints on poses in an online manner to maintain representation that is both robust and adaptive.

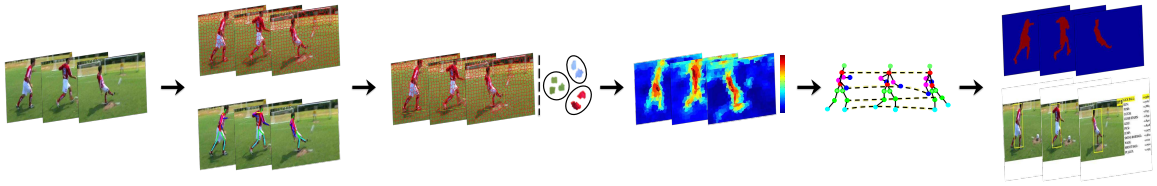
In summary, 1) we address the problem of *Online Action Localization* in a streaming video, 2) by using high-level pose estimation to learn a mid-level superpixel-based foreground model at each time instant. 3) The label and confidences for action segments are *predicted* using dynamic programming on SVM scores trained on partial action clips. Finally, 4) we also introduce an evaluation measure to quantify performance of action prediction and online localization. The rest of the paper is organized as follows. In Sec. 2 we review literature relevant to our approach. Sec. 3 covers the technical details of our approach. We report results in Sec. 4 and conclude with suggestions for future work in Sec. 5.

2. Related Work

Online Action Prediction aims to predict actions from partially observed videos *without* any localization. These methods typically focus on maximum use of temporal, sequential and past information to predict labels. Li and Fu [18] predict human activities by mining sequence patterns, and modeling causal relationships between them. Zhao *et al*. [45] represent the structure of streaming skeletons (poses) by a combination of human-body-part movements and use it to recognize actions in RGB-D. Hoai and De la Torre [8] simulate the sequential arrival of data while training, and train detectors to recognize incomplete events. Similarly, Lan *et al*. [16] propose hierarchical 'movemes' to describe human movements and develop a max-margin learning framework for future action prediction.

Ryoo [26] proposed integral and dynamic bag-of-words for activity prediction. They divide the training and testing videos into small segments and match the segments sequentially using dynamic programming. Kong *et al*. [14] proposed to model temporal dynamics of human actions by explicitly considering all the history of observed features as well as features in smaller temporal segments. Yu *et al*. [43] predict actions using Spatial-Temporal Implicit Shape Model (STISM), which characterizes the space-time structure of the sparse local features extracted from a video. Cao *et al*. [2] perform action prediction by applying sparse coding to derive the activity likelihood at small temporal segments, and later combine the likelihoods for all the segments. In contrast, we perform both action prediction as well as localization in an online manner.

***Offline* Action Localization** has received significant attention in the past few years [41, 38, 9, 5, 13, 35, 11]. The first category of approaches uses either rectangular tubes or cuboid-based representations. Lan *et al*. [17] treated the human position as a latent variable, which is inferred simultaneously while localizing an action. Yuan *et al*. [44] used branch-and-bound with dynamic programming, while Zhou *et al*. [46] used a split-and-merge algorithm to obtain action segments that are then classified with LatentSVM [6]. Oneata *et al*. [22] presented an approximation to Fisher Vectors for tractable action localization. Tran *et al*. [30] used Structured SVM to localize actions with inference performed using Max-Path search method. Ma *et al*. [19] automatically discovered spatio-temporal root and part filters, whereas Tian



(a) Input Stream of Video Frames | (b) Superpixel Extraction and Pose Estimation | (c) Learn Superpixel based Appearance Model | (d) Superpixel based Foreground Likelihood | (e) Pose Refinement | (f) Segment Action with CRF + Action Prediction using SVM

Figure 2. This figure shows the framework of the approach proposed in this paper. (a) Given an input video, (b) we over-segment each frame into superpixels and detect poses using an off-the-shelf method [40]. (c) An appearance model is learned using all the superpixels inside a pose bounding box as positive, and those outside as negative samples. (d) In a new frame, the appearance model is applied on each superpixel of the frame to obtain a foreground likelihood. (e) To handle the issue of visual drift, poses are refined using spatio-temporal smoothness constraints on motion and appearance. (f) Finally, a CRF is used to obtain local action proposals, which are then utilized to predict the action through dynamic programming on SVM scores.

et al. [29] developed Spatio-temporal Deformable Parts Model [6] to detect actions in videos and can handle deformities in parts, both in space and time. Recently, Yu and Yuan [42] proposed a method for generating action proposals obtained by detecting tubes with high actionness scores after non-maximal suppression.

The second category uses either superpixels or supervoxels as the base representations [21, 10]. Jain *et al*. [10] recently proposed a method that extends selective search approach [31] to videos. They merge supervoxels using appearance and motion costs and produce multiple layers of segmentation for each video. Gkioxari and Malik [7] use selective search [31] to generate candidate proposals for video frames, whose spatial and motion Convolutional Neural Network (CNN) features are evaluated using SVMs. The per-frame action detections are then linked temporally for localization. There have been few similar recent methods for quantifying actionness [3, 42], which yield fewer regions of interest in videos. Similar to these methods, our approach can delineate contours of an action, but with the goal of performing prediction and localization in a streaming fashion.

Pose for Action Recognition was used by Maji *et al*. [20], who implicitly captured poses through 'poselet activation vector' and employed those for action recognition in static images. However, such a representation is not useful for detecting an action foreground. Xu *et al*. [39] detect poses through [40] and couple them with independently computed local motion features around the joints for action recognition. Wang *et al*. [33] also extended [40] to videos and represented videos in terms of spatio-temporal configurations of joints to perform action recognition. Raptis and Sigal [24] recognize and detect interactions from videos by modeling poselets as latent variables in a structured SVM formulation. Joint recognition of action and pose estimation in videos was recently proposed by Nie *et al*. [37]. They divide the action into poses, spatio-temporal parts and then parts, and model their inter-relationships through And-Or graphs. Pirsiavash *et al*. [23] predict quality of sports actions by training a regression model from spatio-temporal pose features, to scores from expert judges. Poses were recently used for *offline* action localization by Wang *et al*. [35], who detect actions using a unified approach that discovers action parts using dynamical poselets, and the relations between them. In contrast to these methods, we use pose in conjunction with low-level iDTF features [34] and mid-level superpixels. Moreover, we predict and localize actions in an online manner in partially observed videos.

3. Localizing and Predicting Actions

The proposed approach (Fig. 2) begins by segmenting the testing video frames into superpixels and detecting pose hypotheses within each frame. The features computed for each superpixel are used to learn a superpixel-based appearance model, which distinguishes the foreground from the background. Simultaneously, the conditional probability of pose hypotheses at current time-step (frame) is computed using pose confidences and consistency with poses in previous frames. The superpixel and pose-based foreground probability is used to infer the action location at each frame through Conditional Random Field. The action label is predicted within the localized action bounding box through dynamic programming using scores from Support Vector Machines (SVMs) on short video clips. These SVMs were trained on temporal segments of the training videos. After localizing action at each time-step (frame), we refine poses in a batch of few frames by imposing spatio-temporal consistency. Similarly, the appearance model is updated to avoid visual drift. This process is

repeated for every frame in an online manner (see Fig. 2) and gives action localization and prediction at every frame.

Let $\mathbf{s}_t$ represent superpixels by its centroid in frame t and $\mathbf{p}_t$ represent poses in frame t. Since our goal is to localize the action in each frame, we use $\mathbf{X}_t$ to represent, a sequence of bounding boxes (tube) in a small window of δ frames. Each bounding box is represented by its centroid, width and height. Similarly, let $\mathbf{S}_t$ and $\mathbf{P}_t$ respectively represent all the superpixels and poses within that time window. Given the pose and superpixel-based observations till time t, $\mathbf{S}_{1:t}$ and $\mathbf{P}_{1:t}$, the state estimate $\mathbf{X}_t$ at time t is obtained using the following equation through Bayes Rule:

$$p(\mathbf{X}_t|\mathbf{S}_{1:t},\mathbf{P}_{1:t}) = Z^{-1}p(\mathbf{S}_t|\mathbf{X}_t).p(\mathbf{P}_t|\mathbf{X}_t). \int p(\mathbf{X}_t|\mathbf{X}_{t-1}).p(\mathbf{X}_{t-1}|\mathbf{S}_{1:t-1},\mathbf{P}_{1:t-1})d\mathbf{X}_{t-1}, \quad (1)$$

where Z is the normalization factor, and the state transition model is assumed to be Gaussian distributed, i.e., $p(\mathbf{X}_t|\mathbf{X}_{t-1}) = \mathcal{N}(\mathbf{X}_t;\mathbf{X}_{t-1},\Sigma)$. Eq. 1 accumulates the evidence over time on the superpixels and poses in batch-streaming mode. The state which maximizes the posterior (MAP) estimate in Eq. 1 is selected as the new state. Next, we define the pose and superpixel based foreground likelihoods used for estimating Eq. 1.

3.1. Superpixel-based Foreground Likelihood

Learning an appearance model helps in distinguishing the foreground actions from the background. Given foreground and background superpixels in the previous frames $t-\delta:t-1$, we group them into $k=1\dots K$ clusters. Furthermore, let ξ_k define the ratio of foreground to background superpixels for the kth cluster. Then, the appearance-based foreground score is given by:

$$H_{\text{fg}}(\mathbf{s}_t) = \exp\left(\frac{\|\boldsymbol{\phi}_{\text{color}}(\mathbf{s}_t)-\mathbf{c}_k)\|}{r_k}\right)\cdot\xi_k + \exp\left(\frac{\|\boldsymbol{\phi}_{\text{flow}}\mathbf{s}_t)-\boldsymbol{\mu}_k\|}{\sigma_k}\right), \quad (2)$$

where $\mathbf{c}_k$ is the center, r_k is the radius, $\boldsymbol{\mu}_k$ is the mean optical flow and σ_k is the flow variance for kth cluster.

In Eq. 2, the clusters are updated incrementally at each time-step (frame) to recover from the visual drift using a temporal window of past δ frames. Note that, background pixels within a foreground bounding box are inevitably considered as foreground, and introduce noise during model update. The ξ_k helps to compensate for this issue by quantifying the foreground/background ratio for each cluster. Finally, the superpixel-based foreground likelihood in Eq. 1 is given as: $p(\mathbf{S}_t|\mathbf{X}_t) = \alpha_{\text{fg}}\cdot H_{\text{fg}}(\mathbf{s}_t)$, where α_{fg} is the normalization factor.

3.2. Pose-based Foreground Likelihood

We use a pre-trained pose detector to obtain pose hypotheses in each frame. Each pose $\mathbf{p}_t$ is graphically represented with a tree, given by $T=(\Pi,\Lambda)$. The body joints $\pi\in\Pi$ are based on appearance, that are connected by $\lambda\in\Lambda$ edges capturing deformations. The joint j with its location in pose $\mathbf{p}_t$ is represented by $\boldsymbol{\pi}_t^j$, consisting of its x and y locations. Then, the cost for a particular pose $\mathbf{p}_t$ is the sum of appearance and deformation costs:

$$H_{\text{raw}}(\mathbf{p}_t) = \sum_{j\in\Pi_t}\Psi\left(\boldsymbol{\pi}_t^j\right) + \sum_{(j,j')\in\Lambda_t}\Phi\left(\boldsymbol{\pi}_t^j,\boldsymbol{\pi}_t^{j'}\right), \quad (3)$$

where Ψ and Φ are linear functions of appearance features of pose joints, and their relative joint displacements (deformations) w.r.t each other. Poses are obtained using [40], which optimizes over latent variables that capture different joint locations and pose configurations. Since the pose estimation in individual frames is inherently noisy, and does not take into account the temporal information available in videos, we impose the following smoothness constraints in the previous δ frames to re-evaluate poses in Eq. 3.

Appearance Smoothness of Joints: Since the appearance of a joint is not expected to change drastically in a short window of time, we impose the appearance consistency between superpixels at joint locations:

$$J_{\text{app}}(\mathbf{p}_t) = \sum_{j=1}^{|\Pi_t|}\|H_{\text{fg}}(\hat{\mathbf{s}}_t^j) - H_{\text{fg}}(\hat{\mathbf{s}}_{t-1}^j)\|, \quad (4)$$

where $\hat{\mathbf{s}}_t^j$ is the enclosing superpixel of the joint $\boldsymbol{\pi}_t^j$.

Location Smoothness of Joints: We ensure that joint locations are smooth over time. This is achieved by fitting a spline to each joint on the past δ frames, $\boldsymbol{\gamma}_t^j$. Then the location smoothness cost is given by:

$$J_{\text{loc}}(\mathbf{p}_t) = \sum_{j=1}^{|\Pi_t|}\|\boldsymbol{\gamma}_t^j - \boldsymbol{\pi}_t^j\|. \quad (5)$$

Scale Smoothness of Joints: Let $j_{\text{min}}, j_{\text{max}}$ respectively denote minimum and maximum for $\boldsymbol{\gamma}$, i.e. the vertical dimension of the bounding box circumscribing all the splines fitted on joints. And $j'_{\text{min}}, j'_{\text{max}}$ denote minimum and maximum for joints in actual poses $\boldsymbol{\pi}_t\in\Pi_t$. Then, the scale smoothness cost essentially computes the overlap between vertical dimensions of the two:

$$J_{\text{sc}}(\mathbf{p}_t) = \|(j_{\text{max}} - j_{\text{min}}) - (j'_{\text{max}} - j'_{\text{min}})\|. \quad (6)$$

The cost of a particular pose is defined as its raw cost plus the smoothness costs across time, i.e., $H_{\text{pose}}(\mathbf{p}_t) = H_{\text{raw}}(\mathbf{p}_t) + J_{\text{app}}(\mathbf{p}_t) + J_{\text{loc}}(\mathbf{p}_t) + J_{\text{sc}}(\mathbf{p}_t)$. Similar to Sec. 3.1, we use a temporal window of past δ frames to refine the pose locations. We propose an iterative approach to select poses in the past $t-\delta : t$ frames. Given an initial set of poses, we fit a spline to each joint $\boldsymbol{\pi}_t^j$. Then, our goal is to select a set of poses from $t-\delta$ to t frames, such that the following cost function is minimized:

$$({}^*\mathbf{p}_{t-\delta}, \ldots, {}^*\mathbf{p}_t) = \underset{\mathbf{p}_{t-\delta},\ldots,\mathbf{p}_t}{\arg\min} \sum_{\tau=t-\delta}^{t} \Big(H_{\text{pose}}(\mathbf{p}_\tau) \Big). \quad (7)$$

This function optimizes over pose detection, and the appearance, location and scale smoothness costs of joints (see Fig.2 (e)) by greedily selecting the minimum cost pose in every frame through multiple iterations. Finally, the pose-based foreground likelihood in Eq. 1 is given by $p(\mathbf{P}_t|\mathbf{X}_t) = \exp(\alpha_{\text{pose}} \cdot H_{\text{pose}}(\mathbf{p}_t))$, where α_{pose} is the normalization factor.

3.3. Action Localization using CRF

Once we have the superpixel and pose-based foreground likelihoods, we infer the action segments using a history of δ frames. Although the action location is computed online for every frame, using past δ frames adds robustness to segmentation. We form a graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ with superpixels as nodes connected through *spatial* and *temporal* edges. Let variable a denote the foreground/background label of a superpixel. Then, the objective function of CRF becomes:

$$\begin{aligned} &-\log\big(p(a_{t-\delta}, \ldots, a_t|\mathbf{s}_{t-\delta}, \ldots, \mathbf{s}_t, \mathbf{p}_{t-\delta}, \ldots, \mathbf{p}_t)\big) \\ &= \sum_{\tau=t-\delta}^{t} \Big(\underbrace{\Theta\big(a_\tau|\mathbf{s}_\tau, \mathbf{p}_\tau\big)}_{\text{unary potential}} + \underbrace{\Upsilon\big(a_\tau, a'_\tau|\mathbf{s}_\tau, \mathbf{s}'_\tau\big)}_{\text{spatial smoothness}} \Big) \\ &\quad + \sum_{\tau=t-\delta}^{t-1} \underbrace{\Gamma\big(a_\tau, a'_{\tau+1}|\mathbf{s}_\tau, \mathbf{s}'_{\tau+1}\big)}_{\text{temporal smoothness}}, \quad (8) \end{aligned}$$

where the unary potential, with the associated weights symbolized with α, is given by:

$$\Theta\big(a_\tau|\mathbf{s}_\tau, \mathbf{p}_\tau\big) = \alpha_{\text{fg}} H_{\text{fg}}(\mathbf{s}_\tau) + \alpha_{\text{pose}} H_{\text{pose}}(\mathbf{p}_\tau), \quad (9)$$

and the spatial and temporal binary potentials, with weights β and distance functions d, are given by:

$$\begin{aligned} &\Upsilon\big(a_\tau, a'_\tau|\mathbf{s}_\tau, \mathbf{s}'_\tau\big) \\ &= \beta_{\text{col}} d_{\text{col}}(\mathbf{s}_\tau, \mathbf{s}'_\tau) + \beta_{\text{hof}} d_{\text{hof}}(\mathbf{s}_\tau, \mathbf{s}'_\tau) + \beta_\mu d_\mu(\mathbf{s}_\tau, \mathbf{s}'_\tau) \\ &\quad + \beta_{\text{mb}} d_{\text{mb}}(\mathbf{s}_\tau, \mathbf{s}'_\tau) + \beta_{\text{edge}} d_{\text{edge}}(\mathbf{s}_\tau, \mathbf{s}'_\tau), \quad (10) \end{aligned}$$

and

$$\begin{aligned} \Gamma\big(a_\tau, a'_{\tau-1}|\mathbf{s}_\tau, \mathbf{s}'_{\tau-1}\big) &= \beta_{\text{col}} d_{\text{col}}(\mathbf{s}_\tau, \mathbf{s}'_{\tau-1}) \\ &\quad + \beta_{\text{hof}} d_{\text{hof}}(\mathbf{s}_\tau, \mathbf{s}'_{\tau-1}) + \beta_\mu d_\mu(\mathbf{s}_\tau, \mathbf{s}'_{\tau-1}), \quad (11) \end{aligned}$$

respectively. In Eqs. 10, and 11, $\beta_{\text{col}} d_{\text{col}}(.)$ is the cost of color features in HSI color space, $\beta_{\text{hof}} d_{\text{hof}}(.)$ and $\beta_\mu d_\mu(.)$ compute compatibility between histogram of optical flow, and mean of optical flow magnitude, of two superpixels. Similarly, $\beta_{\text{mb}} d_{\text{mb}}(.)$ and $\beta_{\text{edge}} d_{\text{edge}}(.)$ quantify incompatibility between superpixels with prominent boundaries.

3.4. Action Prediction

Since localization requires predicting the class of an action in every frame of the streaming video, we make online prediction of actions in the tubes localized by our approach. Our aim is to capture the sequential information present in a video. For training, we divide the videos into 1 second temporal segments and train an SVM on each segment $0 \rightarrow 1$ sec, $1 \rightarrow 2$ sec, $\ldots$ of all the videos for that action. Given testing video segments, we apply SVM classification for all the 1 second training segments to each segment of the test video, and then use dynamic programming to accumulate matching confidences of the most similar sequence. At each step of the dynamic programming, the system effectively searches for the best matching segment that maximizes the SVM confidences from past segments. This method is applied independently for each action, and gives the confidence for each action. This shares resemblance to Dynamic Bag-Of-Words approach [26] who used RBF function to compute distance between training and testing segments. Note that the performance of classification for action localization depends on the quality of localized tubes / cuboids, as the classifiers are only evaluated on such video segments. This is in contrast to other action prediction methods [26, 18, 14, 8] which do not spatially localize the actions of interest.

4. Experiments

We evaluate our *online action localization* approach on two challenging datasets: 1) JHMDB and 2) UCF Sports. We provide details for the experimental setup followed by the performance evaluation and analysis of the proposed algorithm.

Features: For each frame of the testing video we extract superpixels using SLIC [1]. This is followed by extraction of color features (HSI) for each superpixel in the frame, as well as improved Dense Trajectory features (iDTF: HOG, HOF, MBH, Traj) [34] within the

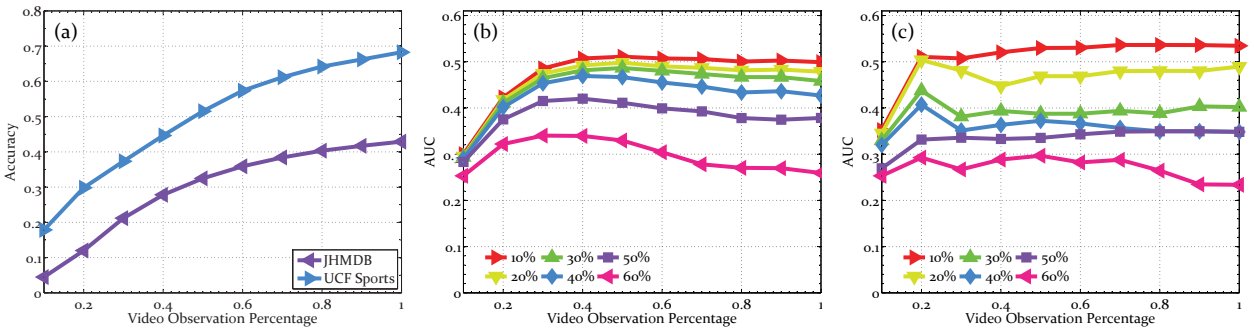


Figure 3. This figure shows action prediction and localization performance as a function of observed video percentage. (a) shows prediction accuracy for JHMDB and UCF Sports datasets; (b) and (c) show localization accuracy for JHMDB and UCF Sports, respectively. Different curves show evaluations at different overlap thresholds: 10% (red), 30% (green) and 60% (pink).

streamed volumes of the video. Each superpixel descriptor has a length of 512 and we set $K = 20$. The pose detections are obtained using [40] and pose features using [12]. We build a vocabulary of 20 words for each pose feature, and represent a pose with 180d vector.

Parameters and Distance Functions: We use Euclidean distance for d_μ, chi-squared distance for d_{hof} and d_{col}, and geodesic distance for d_{mb} and d_{edge}. We normalize the scores used in CRF, therefore, we set absolute values of all the parameters α and β to 1.

Evaluation Metrics: Since the online localization algorithm generates tubes or cuboids with associated confidences, the Receiver Operating Characteristic (ROC) curves are computed at fixed overlap thresholds. Following experimental setup of Lan *et al*. [17], we show ROC @ 20% overlap. Furthermore, Area Under the Curve (AUC) of ROC at various thresholds gives an overall measure of performance.

Inspired from early action recognition and prediction works [26], we also quantify the performance as a function of *Video Observation Percentage*. For this method, the localization and classification for testing videos are sampled at different percentages of observed video $(0, 0.1, 0.2, \ldots, 1)$. The ROC curve is computed at multiple overlap thresholds, and AUC is computed under ROC curves at different thresholds.

Baseline for Online Action Localization: We compare with offline methods which use entire videos to localize actions, and also compute results for a competitive online localization baseline for comparison. For the proposed baseline, we exhaustively generate bounding boxes with overlaps at multiple scales in each frame. These boxes are connected with appearance similarity costs. Over time, the boxes begin to merge into tubes. For temporal window of $\delta = 5$ frames (same as our method), we evaluate each tube with classifiers for all the actions using iDT features.

4.1. Datasets

JHMDB Dataset: The JHMDB [12] dataset is a subset of the larger HMDB51 [15] dataset collected from digitized movies and YouTube videos. It contains 928 videos consisting of 21 action classes. The dataset has annotations for all the body joints and has recently been used for offline action localization [7]. We use a codebook size of 4000 to train SVMs using iDTF features.

UCF Sports Dataset: The UCF Sports [25, 28] dataset consists of 150 videos with 10 action classes. We evaluated our approach using the methodology proposed by Lan *et al*. [17], with a train-test split and intersection-over-union criterion at an overlap of 20%. To train SVMs, we use a codebook size of 1000 on iDTFs.

4.2. Results and Analysis

Action Prediction with Time: The prediction accuracy is evaluated with respect to the percentage of video observed. Fig. 3(a) shows the accuracy against time for JHMDB and UCF Sports datasets. It is evident that predicting the class of an action based on partial observation is very challenging, and the accuracy of correctly predicting the action increases as more information becomes available. An analysis of prediction accuracy per action class is shown in Fig. 4 for (a) JHMDB and (b) UCF Sports datasets. This figure shows that certain actions are more challenging to predict compared to others. For example, the actions *Jump* and *Swing Side* are the most challenging actions when compared to *Golf* and *Kicking*. This is due to the difficulty in correctly estimating pose under high human body articulation.

Since each action has its own predictability, we analyze how early we can predict each action. We arbitrarily set the prediction accuracy to 30% and show the percentage of video observation required for each action of JHMDB and UCF Sports datasets in Table 1. Although we set a reasonable prediction target, certain actions do

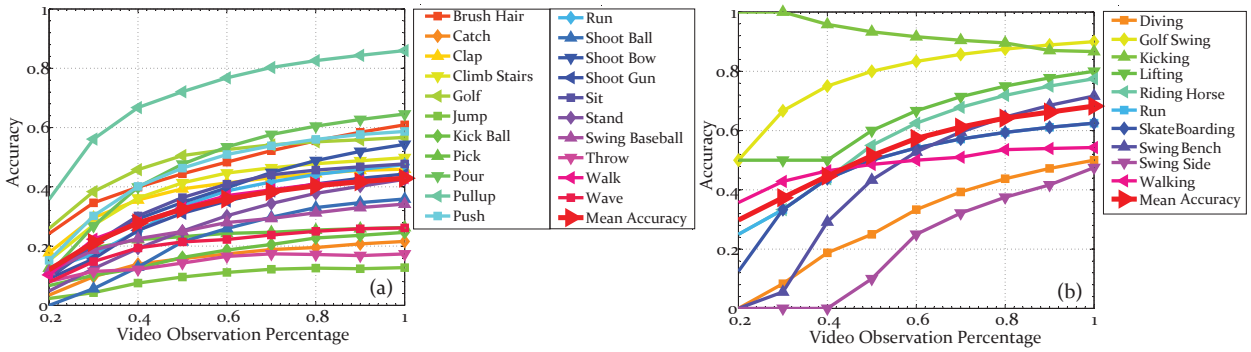


Figure 4. This figure shows per-action prediction accuracy as a function of observed video percentage for (a) JHMDB and (b) UCF Sports datasets.

JHMDB Actions	*Pullup*	*Golf*	*Brush Hair*	*Push*	*Clap*	*Pour*	*Climb Stairs*	*Sit*	*Shoot Bow*	*Walk*	*Run*
Video (%)	*14%*	*23%*	*25%*	*30%*	*31%*	*32%*	*33%*	*40%*	*41%*	*43%*	*45%*
JHMDB Actions	*Shoot Gun*	*Stand*	*Shoot Ball*	*Swing Baseball*	*Pick*	*Wave*	*Kick Ball*	*Catch*	*Throw*	*Jump*	
Video (%)	*48%*	*60%*	*70%*	*70%*	-	-	-	-	-	-	
UCF Sports Actions	*Kicking*	*Lifting*	*Walking*	*Golf Swing*	*Riding Horse*	*Run*	*Skate Boarding*	*Swing Bench*	*Diving*	*Swing Side*	
Video (%)	*1%*	*1%*	*12%*	*16%*	*26%*	*26%*	*28%*	*31%*	*55%*	*67%*	

Table 1. This figure shows the the percentage of video observation required to achieve a prediction accuracy of 30%. Results in the first two rows are from JHMDB, and the last row is from UCF Sports dataset. Actions with missing values indicate that they did not reach a prediction accuracy of 30% until video completion.

not reach such prediction accuracy even until the completion of the video. This shows the challenging nature of online action prediction and localization.

Action Localization with Time: To evaluate online performance, we analyze how the localization performance varies across time by computing accuracy as a function of observed video percentage. Fig. 3(b-c) shows the AUC against the percentage of observed video for different overlap thresholds ($10\% - 60\%$) for (b) JHMDB and (c) UCF Sports. We compute the AUC with time in a cumulative manner such that the accuracy at 50% means localizing an action from start till one-half of the video has been observed. This gives an insight into how the overall localization performance varies as a function of time or observed percentage in testing videos. These graphs show that it is challenging to localize an action at the beginning of the video, since there is not enough discriminative motion observed by the algorithm to distinguish different actions. Furthermore, our approach first learns an appearance model from pose bounding boxes, which are improved and refined as time progresses. This improves the superpixel-based appearance confidence, which then improves the localization, and stabilizes the AUC. The curves also show that the AUC is inversely proportional to the overlap threshold.

There are two interesting observations that can be made from these graphs. First, for the JHMDB dataset in Fig. 3(b), the results improve initially, but then deteriorate in the middle, i.e. when the observation percentage is around 60%. The reason is that most of the articulation and motion happens in the middle of the video. Thus, the segments in the middle are the most difficult to localize, resulting in drop of performance. Second, the curves for UCF Sports in Fig. 3(c) depict a rather unexpected behavior in the beginning, where localization improves and then suddenly worsens at around 15% observation percentage. On closer inspection, we found that this is due to rapid motion in some of the actions, such as *diving* and *swinging (side view)*. For these actions, the initial localization is correct when the actor is stationary, but both actions have very rapid motion in the beginning, which violates the continuity constraints applicable to many other actions. This results in a drop in performance, and since this effect accumulates as the percentage of video observation increases, the online algorithm never attains the peak again for many overlap thresholds despite observing the entire video.

Action Localization with Offline Methods: We also evaluate the performance of our method against existing *offline* state-of-the-art action localization methods. Fig.

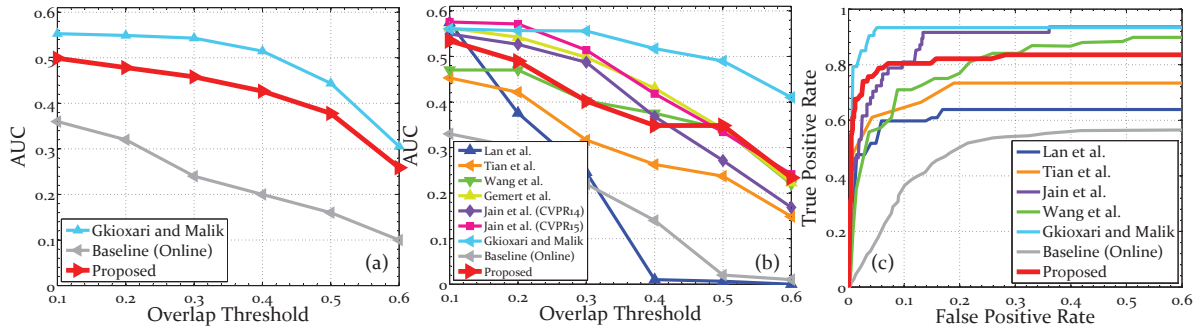


Figure 5. This figure shows localization results of proposed method along with existing methods on JHMDB and UCF Sports datasets. (a) shows AUC curves for JHMDB, while (b) and (c) show AUC and ROC @ 20%, respectively, for UCF Sports dataset. The curves for the proposed method is shown in red, while other offline localization methods including Lan *et al*. [17], Tian *et al*. [29], Wang *et al*. [35], Gemert *et al*. [32], Jain *et al*. [10], Jain *et al*. [11], and Gkioxari *et al*. [7] are shown with different colors, with baseline for online localization in gray.

5(a) shows the results of the proposed method, on JHMDB dataset, in red, and that of [7] in blue. The difference in performance is attributed to the online vs. offline nature of the methods, as well as the use of CNN features by [7]. Furthermore, we outperform a competitive online localization baseline shown in gray. A quantitative comparison on UCF Sports using AUC and ROC @ 20% is shown in Fig. 5(b) and (c) respectively.

Pose Refinement: Pose-based foreground likelihood refines poses in an iterative manner using spatio-temporal smoothness constraints. Our qualitative results in Fig. 6 show the improvement in pose joint locations.

Action Segments: Since we use superpixel segmentation to represent the foreground actor, our approach outputs action segments. Our qualitative results in Fig. 7 show the fine contour of each actor (yellow) along with the ground truth (green). Using superpixels and CRF, we are able to capture the shape deformation of the actors.

5. Conclusion

In this paper, we introduced a new prediction problem of online action localization where the goal is to simultaneously localize and predict actions in an online manner. We presented an approach which uses representations at different granularities - from high-level poses for initialization, mid-level features for generating action tubes, and low-level features such as iDTF for action prediction. We also refine pose estimation in a small batch of frames using spatio-temporal constraints. The localized tubes are obtained using CRF, and classification confidences come from dynamic programming on SVM scores. The intermediate results and ablation study indicate that such an approach is capable of addressing this difficult problem, and performing on par with some of the recent offline action localization methods. For future research, we plan to leverage training data to perform localization and prediction simultaneously by learning costs for superpixel merging for different actions.

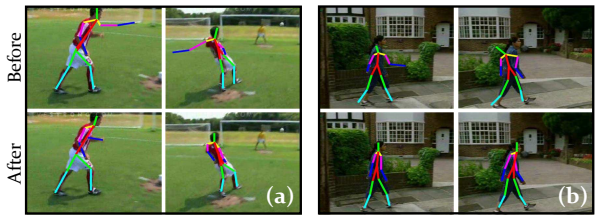


Figure 6. This figure shows qualitative results for pose refinement. Results show a comparison of raw poses (top row) and refined poses (bottom row) for (a) Kicking and (b) Walking.

Figure 7. This figure shows qualitative results of the proposed approach, where each action segment is shown with yellow contour and ground truth with green bounding box. Results in the top three rows are from JHMDB, and the bottom three rows are from UCF Sports datasets.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI*, 34(11), 2012. 5

[2] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. M. Siskind, and S. Wang. Recognize human activities from partially observed videos. In *CVPR*, 2013. 2

[3] W. Chen, C. Xiong, R. Xu, and J. J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *CVPR*, 2014. 3

[4] A. Dehghan, H. Idrees, and M. Shah. Improving semantic concept detection through the dictionary of visually-distinct elements. In *CVPR*, 2014. 1

[5] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*. 2012. 2

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9), 2010. 2, 3

[7] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*. 2015. 3, 6, 8

[8] M. Hoai and F. De la Torre. Max-margin early event detectors. *IJCV*, 107(2), 2014. 2, 5

[9] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang. Action detection in complex scenes with spatial and temporal ambiguities. In *ICCV*, 2009. 2

[10] M. Jain, J. Gemert, H. Jegou, P. Bouthemy, and C. Snoek. Action localization with tubelets from motion. In *CVPR*, 2014. 1, 3, 8

[11] M. Jain, J. C. van Gemert, and C. G. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, 2015. 2, 8

[12] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013. 6

[13] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007. 2

[14] Y. Kong, D. Kit, and Y. Fu. A discriminative model with multiple temporal scales for action prediction. In *ECCV*. 2014. 1, 2, 5

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 6

[16] T. Lan, T.-C. Chen, and S. Savarese. A hierarchical representation for future action prediction. In *ECCV*. 2014. 1, 2

[17] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011. 1, 2, 6, 8

[18] K. Li and Y. Fu. Prediction of human activity by discovering temporal sequence patterns. *IEEE TPAMI*, 36(8), 2014. 1, 2, 5

[19] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff. Action recognition and localization by hierarchical space-time segments. In *ICCV*, 2013. 2

[20] S. Majiwa, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 3

[21] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection proposals. In *ECCV*. 2014. 1, 3

[22] D. Oneata, J. Verbeek, and C. Schmid. Efficient action localization with approximately normalized fisher vectors. In *CVPR*, 2014. 1, 2

[23] H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the quality of actions. In *ECCV*. 2014. 3

[24] M. Raptis and L. Sigal. Poselet key-framing: A model for human activity recognition. In *CVPR*, 2013. 3

[25] M. Rodriguez, A. Javed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 6

[26] M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011. 1, 2, 5, 6

[27] K. Soomro, H. Idrees, and M. Shah. Action localization in videos through context walk. In *ICCV*, 2015. 1

[28] K. Soomro and A. R. Zamir. Action recognition in realistic sports videos. In *Computer Vision in Sports*, pages 181–208. Springer, 2014. 6

[29] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013. 1, 3, 8

[30] D. Tran and J. Yuan. Max-margin structured output regression for spatio-temporal action localization. In *NIPS*, 2012. 2

[31] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 104(2), 2013. 3

[32] J. C. van Gemert, M. Jain, E. Gati, and C. G. Snoek. Apt: Action localization psroposals from dense trajectories. In *BMVC*, volume 2, page 4, 2015. 8

[33] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. In *CVPR*, 2013. 3

[34] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 2, 3, 5

[35] L. Wang, Y. Qiao, and X. Tang. Video action detection with relational dynamic-poselets. In *ECCV*. 2014. 1, 2, 3, 8

[36] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. *arXiv preprint arXiv:1505.04868*, 2015. 2

[37] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu. Joint action recognition and pose estimation from video. In *CVPR*, 2015. 3

[38] Y. Xie, H. Chang, Z. Li, L. Liang, X. Chen, and D. Zhao. A unified framework for locating and recognizing human actions. In *CVPR*, 2011. 2

[39] R. Xu, P. Agarwal, S. Kumar, V. N. Krovi, and J. J. Corso. Combining skeletal pose with local motion for human activity recognition. In *Articulated Motion and Deformable Objects*. 2012. 3

[40] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 3, 4, 6

[41] G. Yu, N. A. Goussies, J. Yuan, and Z. Liu. Fast action detection via discriminative random forest voting and top-k subvolume search. *IEEE Transactions on Multimedia*, 13(3), 2011. 1, 2

[42] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *CVPR*, 2015. 3

[43] G. Yu, J. Yuan, and Z. Liu. Predicting human activities using spatio-temporal structure of interest points. In *ACM MM*, 2012. 2

[44] J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. *IEEE TPAMI*, 33(9), 2011. 2

[45] X. Zhao, X. Li, C. Pang, X. Zhu, and Q. Z. Sheng. Online human gesture recognition from motion data streams. In *ACM MM*, 2013. 2

[46] Z. Zhou, F. Shi, and W. Wu. Learning spatial and temporal extents of human actions for action detection. *IEEE Transactions on Multimedia*, 17(4), 2015. 2